Models optimized for real-world tasks reveal the necessity of precise temporal coding in hearing

Mark R. Saddler[1,2,3], Josh H. McDermott[1,2,3,4]

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
[2]McGovern Institute for Brain Research, MIT, Cambridge, MA, USA
[3]Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA
[4]Program in Speech and Hearing Biosciences and Technology, Harvard, Cambridge, MA, USA

ABSTRACT

Neurons encode information in the timing of their spikes in addition to their firing rates. Spike timing is particularly precise in the auditory nerve, where action potentials phase lock to sound with sub-millisecond precision, but its behavioral relevance is uncertain. To investigate the role of this temporal coding, we optimized machine learning models to perform real-world hearing tasks with simulated cochlear input. We asked how precise auditory nerve spike timing needed to be to reproduce human behavior. Models with high-fidelity phase locking exhibited more human-like sound localization and speech perception than models without, consistent with an essential role in human hearing. Degrading phase locking produced task-dependent effects, revealing how the use of fine-grained temporal information reflects both ecological task demands and neural implementation constraints. The results link neural coding to perception and clarify conditions in which prostheses that fail to restore high-fidelity temporal coding could in principle restore near-normal hearing.

INTRODUCTION

Sensory systems encode information about the environment in the spiking activity of neurons. Decades of experiments have clarified how stimulus properties are represented at different stages of neural processing, but less is known about how this information gives rise to complex human behavior.

In perceptual science, ideal observer models have long been used to analyze which features of a neural code contribute to behavior[1–3]. An ideal observer derives the statistically optimal solution to a perceptual task given the information available at some stage of neural processing[4]. Since evolutionary pressures drive biological perceptual systems in the direction of optimal performance for tasks that are important in the natural environment, comparisons of an organism's behavior to that of optimal task solutions under candidate biological constraints (e.g., a type of neural code) can reveal whether the organism is also operating under those constraints. This approach has produced rigorous computational accounts of some aspects of vision[5–9] and hearing[2,10–14]. However, ideal observers are limited to tasks and stimuli for which provably optimal solutions can be derived (i.e., for which probability distributions of the generative parameters can be specified), precluding most real-world behaviors. Because real-world stimuli and tasks are the ones that biological systems are likely to have been optimized for, ideal observers have had limited applicability in domains where they might otherwise be most useful.

Here, we propose machine learning as an alternative approach to link neural coding to behavior. Contemporary machine learning models are expressive functions that can be optimized to perform real-world tasks with natural stimuli, "learning" solutions from empirical distributions of stimuli and labels rather than mathematical descriptions of a task. Although not provably optimal, such a model can reveal the characteristics of a system optimized for a problem under particular constraints, and thus offer an alternative to the classic ideal observer approach for real-world perception problems that can only be specified empirically. Previous work has shown that human-like behavior can emerge in deep artificial neural networks optimized for natural tasks[15–21], consistent with the idea that humans are shaped by optimization for such tasks. We propose that comparing the behavior of models optimized to perform tasks using different neural representations can reveal the aspects of neural coding necessary for human behavior, and the ecological pressures that drive their use.

Here, we apply this general approach to the problem of temporal coding. Neurons transmit information in the precise timing of their spikes[22] in addition to their time-averaged firing rates. Temporal coding has been identified across multiple sensory modalities[23–27], but spike timing is arguably most precise in the auditory nerve, where action potentials align to the temporal structure of sound with sub-millisecond precision. This precise spike timing plausibly helps to encode the "fine structure" of a sound waveform (i.e., individual pressure oscillations; Fig. 1a). Mammalian auditory nerve fibers phase lock to sound frequencies as high as 3 to 5 kHz[28–30], such that the auditory system in principle has access to this information from the outset. However, whether this information is actually used by the brain remains among the most debated issues in hearing science[31–33].

The one aspect of hearing widely believed to rely on high-fidelity phase locking is sound localization, which depends in part on microsecond-level timing differences between the two ears. Neural circuits for extracting these timing differences from stimulus fine structure are found in many non-human animals. However, for other aspects of hearing there is no consensus.

The issue has remained unresolved for several reasons. First, there is no conclusive evidence for monaural circuits that could extract the information in spike timing. The precision of temporal coding degrades with each synapse along the ascending auditory pathway[34,35], such that physiological mechanisms for extracting information from high-frequency phase locking are likely to be situated early.

Yet despite considerable effort, no such mechanisms for extracting phase locking monaurally have been discovered[32,33]. Second, causal manipulations of phase locking are impractical due to the difficulty of manipulating the nerve in vivo, and because non-human animals do not exhibit many human auditory behaviors. Third, attempts to address the issue psychophysically have been inconclusive[36,33], despite widespread proposals that phase locking is critical for hearing in noise[37–41]. The issue is important to resolve for the design of auditory prosthetics, which at present generally do not reproduce the phase locking seen in the normal ear, and for understanding how different forms of hearing loss (some of which are argued to affect the fidelity of temporal coding[42]) affect behavior.

In addition to not knowing whether and when temporal coding is used by the auditory system, it has also remained unclear why it would or would not be used. Even for sound localization, this remains unsettled, as the upper frequency limit of interaural time difference judgments is lower than the limit of phase locking in the nerve[43–45], for reasons that are not well understood. This question has remained difficult to answer because until recently it was infeasible to model real-world auditory behavior, leaving it unclear to what extent precise timing in the input was computationally important for audition. If phase locking is not needed to obtain good performance in natural auditory tasks, and/or if the circuits to extract it monaurally are prohibitively expensive to implement, it might be discarded by the downstream auditory system.

Classical ideal observers have been applied to aspects of this problem, but were restricted to simple tasks with artificial stimuli (e.g., discrimination of single frequencies[2,12]). The resulting models generally overestimate human performance[2], plausibly because human perceptual systems are not optimized for simple tasks with artificial stimuli. Our approach was to instead investigate the issue in models optimized for ecological tasks, by training machine learning systems to perform such tasks using simulated auditory nerve representations as input. To ask whether the information encoded via peripheral spike timing is necessary to account for behavior, we separately optimized models with altered nerve phase locking and compared the models' behavior to that of human listeners (Fig. 1b). The results provide new evidence for the importance of high-fidelity temporal coding in perception, and provide a way to understand why it is used, by showing where it is critical for real-world task performance.

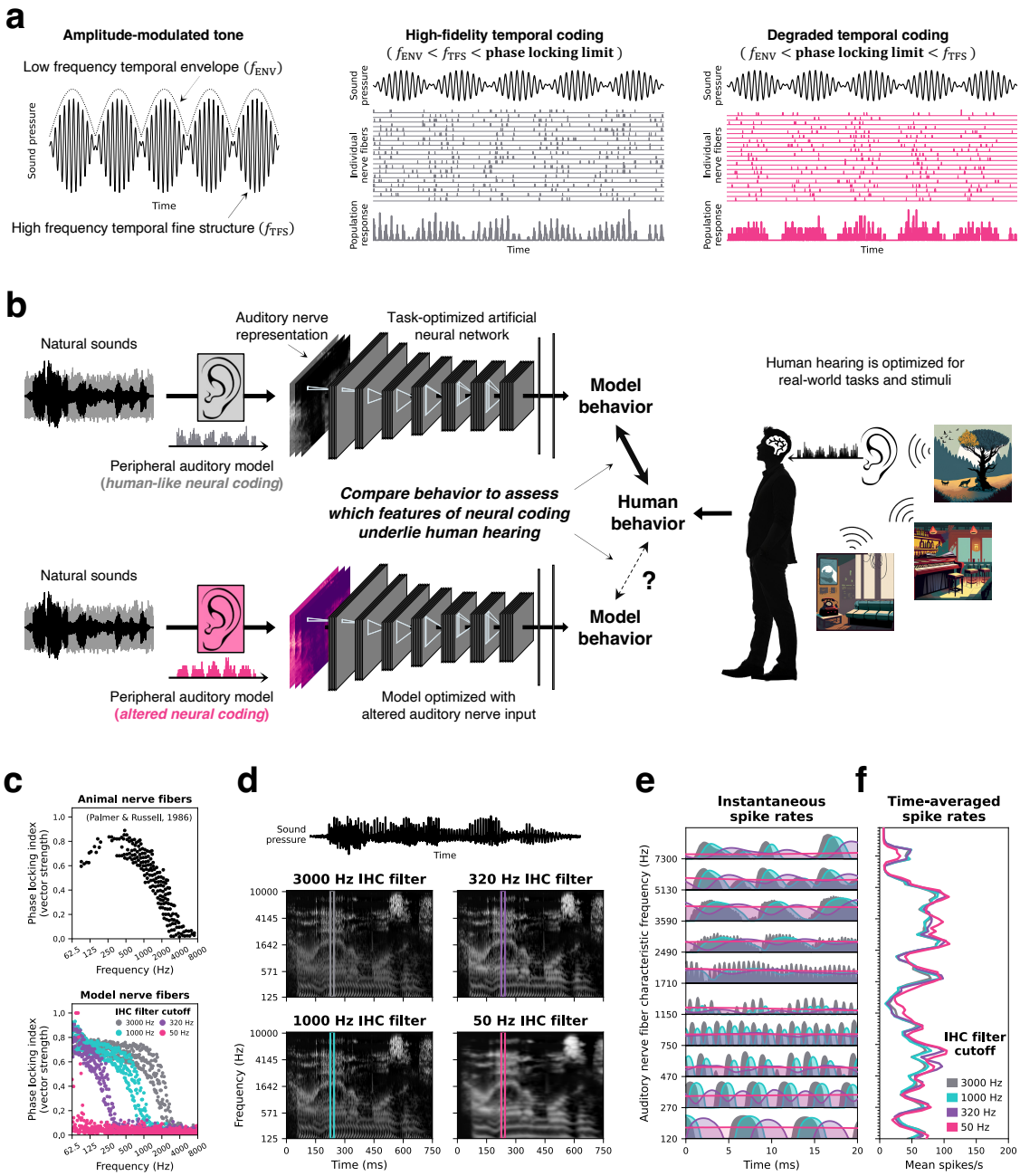


Fig. 1 | Overview of approach. a. Sound waveforms carry information in their amplitude envelope as well as their individual pressure oscillations (the "temporal fine structure" or "TFS"). The envelope and fine structure are encoded with phase-locked spike timing in the auditory nerve. As temporal coding is degraded, auditory nerve spikes no longer phase lock to the fine structure, encoding only the slower envelope in the short-term rate of firing. **b.** Schematic of the approach. Human auditory behavior is shaped by the ears and the acoustic environment. Models optimized to perform naturalistic tasks might reproduce human-like behavior if optimized for the auditory nerve information used by the human auditory system. **c.** Top: The strength of phase locking as a function of frequency, measured in the auditory nerve fibers of guinea pigs. Data is re-plotted from ref[30]. Bottom: The roll-off in phase locking strength is determined by the low-pass filter characteristics of the inner hair cell. Manipulating the hair cell low-pass filter cutoff in model auditory nerve fibers changes the upper frequency limit of phase locking. The 3000 Hz cutoff best approximates the guinea pig data and is often proposed to approximately match the human auditory nerve. **d.** Simulated auditory nerve representations of the same speech waveform with four different configurations of the auditory nerve model. Configurations differed in the inner hair cell low-pass filter cutoff. The 3000 Hz cutoff is commonly used to model the human auditory system. **e.** Instantaneous firing rates from example auditory nerve fibers illustrate the degradation of precise spike timing as the phase locking limit is lowered. Note the rapid oscillations in firing that are present for higher phase locking limits, but absent when the limit is lowered. **f.** Time-averaged firing rates (across the 25ms window depicted in **e**) illustrate that lowering the phase locking limit does not disrupt "place" cues in the overall pattern of excitation across the cochlear frequency axis.

RESULTS

Auditory nerve model stage

We hard-coded the model input representation to approximate the information the ear sends to the brain. We used a phenomenological model[46] of the auditory periphery to simulate instantaneous firing rate responses of a population of auditory nerve fibers whose frequency tuning and sensitivity was intended to match that of the human ear. We simulated the 3 canonical auditory nerve fiber types found in mammals, which have different spiking thresholds and spontaneous activity[47]. High-spontaneous-rate fibers have low thresholds but narrow dynamic ranges, such that their firing rates saturate at conversational sound levels. Medium- and low-spontaneous-rate fibers have higher thresholds and broader dynamic ranges but are less numerous in the ear. The resulting frequency-by-time-by-fiber-type array of instantaneous firing rates was then converted to an array of sampled spike counts, representing the population response of 32000 individual auditory nerve fibers per ear (60% high-spontaneous-rate, 25% medium-spontaneous-rate, and 15% low-spontaneous-rate), which served as the input representation to the networks. To our knowledge, our models are the first to perform naturalistic tasks using a near-realistic representation of the information from a sensory receptor organ.

Temporal coding manipulation

The fidelity of temporal coding in the mammalian ear is limited by ion channels in the hair cell membrane[30] and the hair-cell-to-nerve-fiber synapse[48], which act as low-pass filters. The upper limit of phase locking was altered in silico by changing the cutoff frequency of the low-pass filter governing the inner hair cell potential in the peripheral model. We optimized machine models with different cutoffs to ask whether phase locking was necessary to obtain human-like behavior.

In one training condition, this low-pass cutoff was set to a default value of 3000 Hz, which causes phase locking on par with that seen in electrophysiological recordings from non-human animals (Fig. 1c). This upper limit is presumed to be shared by humans[49–52] but is not directly measurable. To investigate the behavioral relevance of temporal coding, we also trained models with each of three lower cutoff values: 1000 Hz, 320 Hz, and 50 Hz. Lowering this cutoff degrades the fidelity of temporal coding, progressively blurring the auditory nerve representation along the time-axis (Fig. 1d and e). The lowest cutoff eliminates essentially all phase locking to temporal fine structure in natural sounds. However, as expected, the manipulation had very little effect on the pattern of firing rates across the cochlear frequency axis (Fig. 1f): nerve fibers with low phase locking limits encode high frequency sounds in their firing rates, just not with precise spike timing. We separately optimized neural networks operating on auditory nerve representations with these four different cutoff frequencies.

Artificial neural network model stages and training

The neural network portion of each model consisted of a feedforward series of stages instantiating linear convolution, nonlinear rectification, normalization, and pooling. The parameters of these model stages were optimized to perform auditory tasks via supervised machine learning. Each task was operationalized as a classification problem with a single ground-truth label per stimulus.

The performance of a neural network depends on both the weights (that are optimized via gradient descent for training task performance) and the hyperparameters that define the network architecture (e.g., the number of layers and the size and shape of convolutional filter kernels). To ensure that these hyperparameters were also optimized for the tasks, we used the top 10 best-performing network architectures previously identified in large-scale random architecture searches conducted for each task (Supplementary Tables 1-2)[19,20,53]. Results for each task and cochlear model configuration are

presented as the average of these 10 network architectures, allowing us to provide uncertainty estimates and marginalize across the idiosyncrasies of any single network architecture.

Our approach relied on optimizing models for "natural" tasks, on the grounds that these are likely to have shaped the nervous system's strategies. As such, we define natural tasks to be those that humans perform in their daily lives and that have likely been important for survival: recognizing and localizing everyday sounds in everyday conditions. We contrast such tasks with those often used in laboratory experiments, where both the behavioral judgment and stimuli can be artificial (e.g. discriminating synthetic tones or noise signals). Accordingly, training stimuli were compiled from large-scale corpora of natural sounds (speech and recordings of auditory scenes) and were meant to approximate the "auditory diet" that likely shaped biological hearing systems over the course of evolution and development.

Models were optimized to perform three different auditory tasks: sound localization, voice recognition, and word recognition. For each task, we separately trained models with four different auditory nerve phase locking limits and then compared their behavior to that of humans. The models featured here build on previous models of human sound localization[20] and speech recognition[17,54] but were improved in several respects to enable a strong test of the importance of temporal coding. Specifically, they operated on more realistic input representations (incorporating spikes and multiple types of auditory nerve fibers), were trained on more realistic datasets, and were evaluated with an expanded set of psychoacoustic experiments.

Logic of approach and aggregate results

We begin by outlining the logic of the approach along with a summary of the results that illustrates how the overall results relate to this logic and the conclusions that follow from it. We then present the results of individual experiments in each of the three task domains, which illustrate the specific effects that underlie the overall results.

Effect of temporal coding on naturalistic task performance

For each of the three tasks, we first evaluated models tested on naturalistic stimuli in noise, asking whether phase locking is necessary for good performance. Because lowering the phase locking limit removes information from the model's input, there are two main qualitative possibilities. Performance could worsen at lower phase locking limits, which would provide evidence that fine-grained temporal information from phase locking is beneficial for the task, and thus might have driven its role in perception. Alternatively, performance could remain similar across phase locking limits. This result would indicate that fine-grained temporal information is not needed for the task in question.

It was a priori likely that high-fidelity phase locking would matter to some extent for sound localization, where microsecond-level timing differences between the two ears can plausibly only be transduced via spike timing, and where corresponding neural circuitry has been documented. However, it was unclear what to expect for word and voice recognition.

Figure 2 shows the effect of the phase locking cutoff on overall performance for each of the three tasks in noise (solid lines; left y-axes). As expected, sound localization (Fig. 2a) was worse for lower cutoffs, with mean absolute localization error increasing 6.4° as the cutoff was lowered from 3000 to 50 Hz ($p < 0.001$, evaluated by bootstrapping across 10 neural network architectures). However, the effect was driven by cutoffs below 1000 Hz. Voice recognition (Fig. 2b) shows a comparably large effect of phase locking, with accuracy dropping by 10.5% between the 3000 and 50 Hz conditions ($p < 0.001$). By contrast, the effect on word recognition (Fig. 2c) was modest, with accuracy dropping only 2.1% across

conditions (this difference was statistically significant as results across model architectures were highly consistent but is small in absolute terms; $p < 0.001$). These results indicate that high-fidelity temporal coding aids localization and voice recognition in natural conditions but is less critical for word recognition.

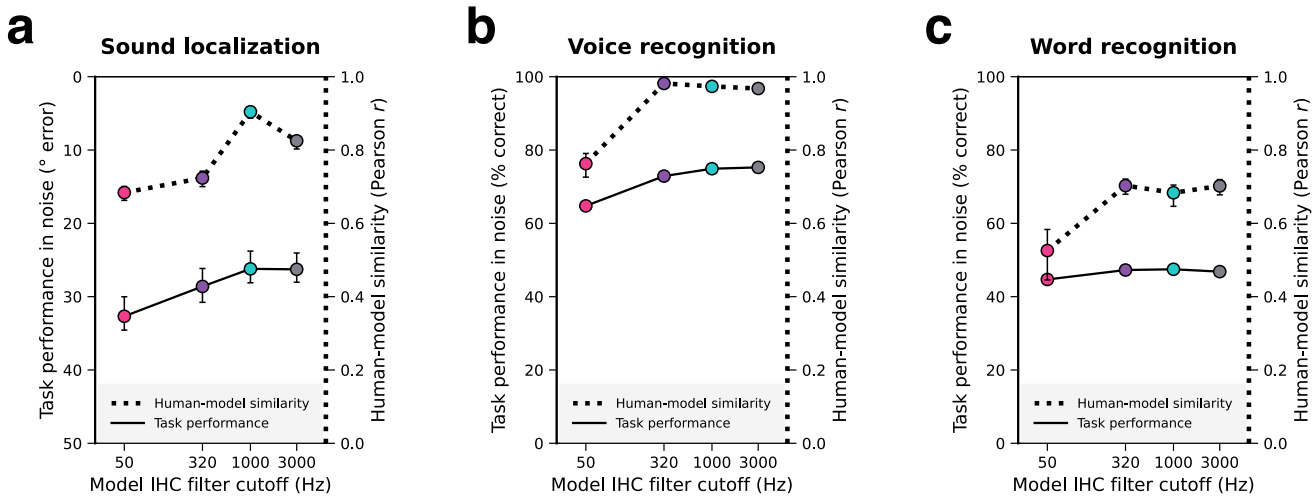


Fig. 2 | Models with access to phase-locked spike timing have better and more human-like hearing. Each panel corresponds to a different task and summarizes the effect of auditory nerve phase locking limit on naturalistic model task performance and overall human-model behavioral similarity. Naturalistic task performance is quantified as a single number averaged across noise conditions shown in later figures (left y-axes; solid lines). Overall human-model behavioral similarity is quantified as the Pearson correlation between analogous human and model data points, averaged across all experiments for each model task (right y-axes; dotted lines). Individual experiments are described in subsequent results sections and figures. Error bars indicate 95% confidence intervals bootstrapped across 10 network architectures for each model. **a.** Sound localization. The left y-axis plots mean absolute error for the sound localization model and is inverted so that better model performance corresponds to higher positions on the y-axis. **b.** Voice recognition. Here and in **c,** the left y-axes plot percent correct for the model when tested on speech in noise. **c.** Word recognition.

Effect of temporal coding on human-model behavioral similarity

For each task we then simulated a battery of psychoacoustic experiments measuring the effect of different cues on perception (described in detail in subsequent sections). We asked whether phase locking was necessary for a model to exhibit human-like behavior as assessed via the pattern of performance across the battery of experiments. The most diagnostic result would be that one of the phase locking limits produces more human-like behavior than the others. Such a result would indicate that phase locking up to that limit contributes to that behavior (and thus must be extracted by the auditory system). The maximally human-similar model could in principle occur for a low phase locking limit, if high-frequency phase locking is discarded by the auditory system for that task.

Figure 2 also shows human-model behavioral similarity in each of the three task domains for each phase locking cutoff (dotted lines; right y-axes). The results implicate phase locking in all three types of behavior, but to different extents. For sound localization the 1000 Hz cutoff produces most human-like behavior (as evaluated by the correlation between human and model results; see Supplementary Fig. 1 for comparable results using root-mean-squared error as the measure of similarity). This result implicates phase locking up to but not above 1000 Hz in human sound localization. By contrast, for both voice and word recognition, the 50 Hz cutoff produced a worse match to human behavior, with the three highest cutoffs producing comparably human-like behavior. This result provides evidence that phase locking above 50 Hz is used in human perception in these domains, with no evidence that phase locking above 320 Hz is used.

Insight into why phase locking is used when it is may be obtained by comparing the dotted and solid lines in Figure 2. For sound localization (Fig. 2a) and voice recognition (Fig. 2b), phase locking is implicated in human-like model behavior while also being necessary for good performance in natural conditions. This combination of results suggests that phase locking is used by the brain because it is advantageous for those tasks. Moreover, the finding that localization performance does not continue to improve above 1000 Hz (Fig. 2a, solid line) provides a normative explanation for why phase locking above 1000 Hz does not appear to influence human perception (Fig. 2a, dotted line). If phase-locked spike-timing information above 1000 Hz does not aid localization in naturalistic conditions, then there would be little evolutionary pressure to extract it. Other species with smaller heads might require higher-fidelity temporal coding of time differences for good localization performance; see Discussion.

By contrast, for word recognition, the human-model similarity results implicate phase locking in human behavior (Fig. 2c, dotted line) despite it providing little performance benefit in natural conditions (Fig. 2c, solid line). This pattern of results suggests that the use of phase locking in word recognition is driven by its importance for other tasks. In the following sections, we consider each task in turn, showing the effect of phase locking on performance at different noise levels, and on a large set of psychophysical assays.

Model optimization -- sound localization

To assess sound localization behavior, models were tasked with reporting the location of target sound sources in naturalistic auditory scenes rendered as binaural audio using a virtual acoustic room and head simulator (Fig. 3a). There were 1.8 million training scenes, each consisting of a target source rendered at a single location (in one of 1800 rooms varying in size, shape, and reverberation) in the presence of spatially diffuse texture-like background noise. The model classified the azimuth and elevation of the target source relative to the simulated listener's head (504-way classification into 5° azimuth by 10° elevation bins). The model operated on auditory nerve responses from the simulated listener's left and right ears, and thus had access to the same monaural and binaural cues as a human listener in the same scene (Fig. 3b). Models optimized for this task with access to high-fidelity temporal coding in the peripheral representation have previously been shown to replicate characteristics of human sound localization[20], including the frequency-dependent use of interaural time and level differences for azimuth judgments[55] and the use of ear-specific spectral cues for elevation judgments[56,57]. However, it was unclear what would happen if models were optimized without high-fidelity temporal coding. Specifically, it was unclear how impaired temporal coding would affect the use of different localization cues, and whether the upper limit of time difference encoding evident in humans could be explained by what is needed for natural task performance.

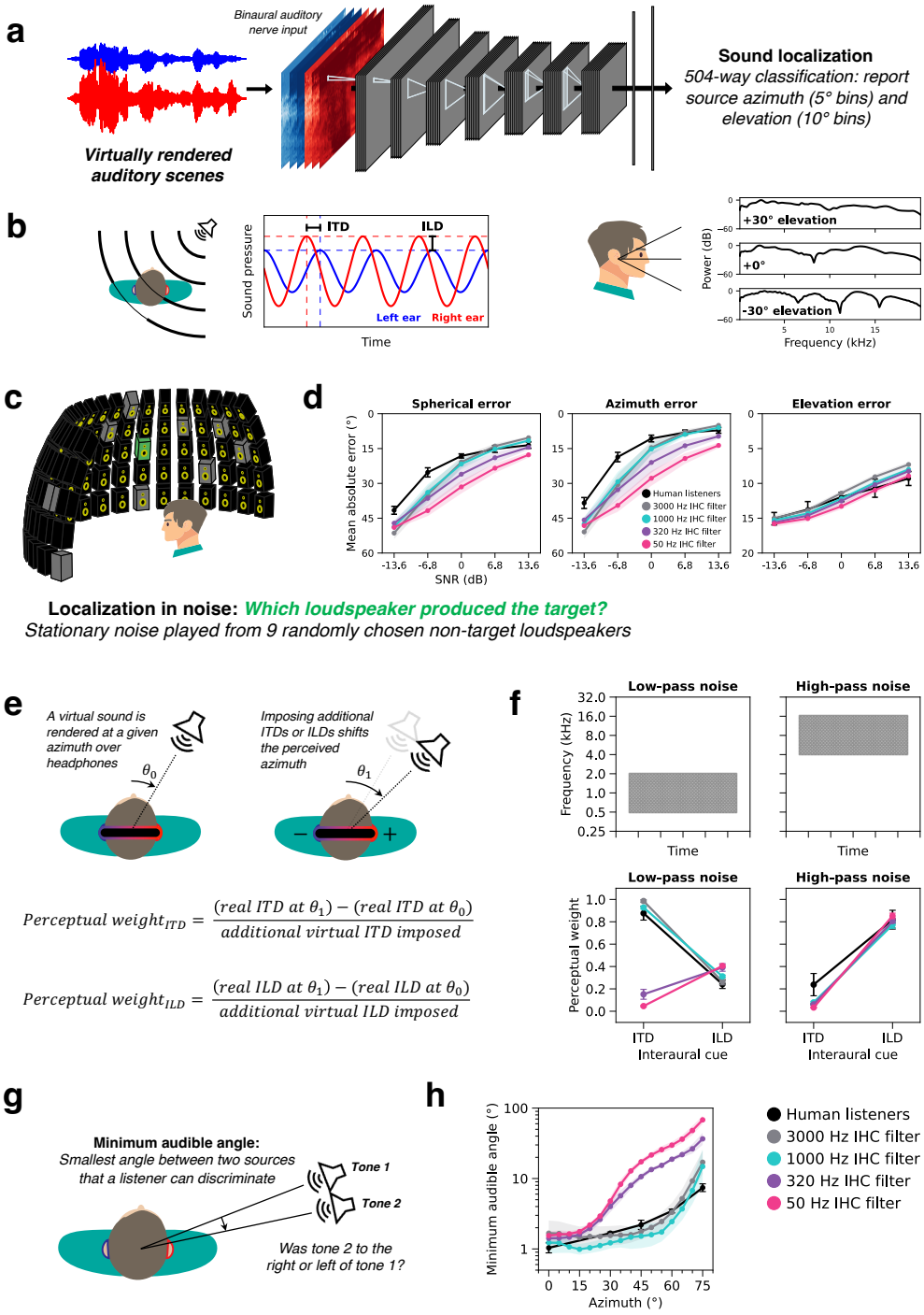


Fig. 3 | Sound localization is impaired in models with degraded auditory nerve spike timing. a. Localization model schematic. Deep artificial neural networks optimized for sound localization operated on binaural auditory nerve representations of virtually rendered auditory scenes. Nerve representations from the left and right ear were supplied as distinct channels to the first neural network stage. **b.** Sound localization cues available to human listeners. Left: interaural time and level differences (ITDs and ILDs) are schematized with pure tones recorded at the left and right ear. Right: spectral differences in the anatomical transfer function provide a monaural cue to elevation. **c.** Schematic of the sound localization in noise experiment. **d.** Mean absolute error for humans and models localizing natural sounds in noise are plotted as a function of SNR. The three axes separately plot spherical, azimuth, and elevation errors. **e.** Schematic of the ITD / ILD cue weighting experiment. The perceptual weights measure the extent to which added ITDs or ILDs shift the perceived azimuth of a virtual sound presented over headphones. **f.** ITD and ILD perceptual weights measured with low-pass and high-pass noise from humans and models. **g.** Schematic of minimum audible angle experiment. **h.** Minimum audible angles plotted as a function of azimuth for human and model listeners. Model error bars always indicate ±2 standard errors of the mean across 10 network architectures per phase locking condition. In **d** and **f**, human error bars indicate ±2 standard errors of the mean across participants. In **h**, human error bars indicate ±2 standard errors from 1 listener averaged across 4 different pure tone frequencies. Human data in **f** and **h** are re-plotted from the original studies[55,58].

Degraded temporal coding impairs sound localization

We compared human and model sound localization accuracy for a set of 460 natural sounds presented in different levels of background noise (Fig. 3c). Humans were asked to report which of 95 loudspeakers (in a 19-by-5 array spanning -90° to 90° azimuth and 0° to 40° elevation in steps of 10°) produced the target sound. This task was intended to tap into some of the challenges (background noise, many possible locations) that accompany localization in real-world settings. On each trial threshold equalizing noise[59] played diffusely from 9 other randomly selected speaker locations. Models performed the same task in a virtual rendering of the loudspeaker array room. Overall task performance was quantified with mean absolute localization error as a function of SNR. Although human listeners outperformed all models at the lowest SNRs (plausibly because the models occasionally report the location of the noise rather than the target), models with access to high-frequency phase locking produced the best match to human behavior (Fig. 3d). Models with 3000 and 1000 Hz phase locking limits exhibited near-human-level robustness to noise, while models with degraded temporal coding made progressively larger localization errors as the phase locking limit was lowered. Degraded temporal coding impaired localization performance in both azimuth and elevation though the effect was larger for azimuth (Fig. 3d, middle vs. right). These results confirm that precise spike timing is important for localizing natural sounds, particularly in noisy environments.

Auditory nerve phase locking is critical for ITD-based sound localization

Biological sound localization relies on three main cues. Small time and level differences between the sound at the ears provide cues to a source's location in the azimuthal plane (Fig. 3b, left). In addition, before impinging on the ear drum, a sound waveform is altered by the pinna, head, and torso, which boost some frequencies and attenuate others. This anatomical filtering is direction-specific, providing a third cue to a source's location (Fig. 3b, right). Humans rely on these "spectral" cues to judge elevation[60,61]. To investigate the contribution of temporal coding to each of these localization cues, we simulated a set of classic psychoacoustic experiments on the models.

Humans rely more on interaural time differences (ITDs) at low frequencies and interaural level differences (ILDs) at high frequencies[62]. One such demonstration comes from measurements of human sensitivity to interaural cue manipulations with virtual sounds[55] (Fig. 3e). In the original experiment, sounds were rendered at different azimuths using a virtual acoustic simulator. Interaural cue sensitivity was inferred from how much a sound's perceived location appeared to shift as additional ITDs or ILDs were added to the binaural waveforms. Shifts in perceived azimuth were mapped back to units of ITD or ILD (as the ITD or ILD change corresponding to an actual shift in azimuth by the same amount), allowing interaural cue sensitivity to be quantified as a dimensionless weight: the slope of the response cue value relative to the imposed cue value. For low-frequency sounds, the ITD weight in humans is much larger than the ILD weight. The reverse is true for high-frequency sounds.

Although the encoding of ITDs is thought to make use of phase locking, it was a priori not entirely clear what to expect from the models with altered phase locking limits. The ITDs of natural sounds are present in amplitude envelopes in addition to the fine structure within frequency channels[63–65]. Because the amplitude envelope is lowpass, interaural envelope delays should in principle be detectable even if the effective sampling rate of cochlear transduction is lowered via the phase locking limit, and could potentially produce ITD sensitivity even without high-fidelity phase locking.

To investigate the contribution of phase locking to this frequency-specific cue dependence, we simulated this experiment on our models (Fig. 3f). Models with high phase locking limits replicated human behavior, exhibiting high ITD sensitivity only for low frequencies and high ILD sensitivity only for high frequencies. Models with degraded temporal coding (320 and 50 Hz phase locking limits)

deviated from human behavior, progressively losing ITD sensitivity at all frequencies and gaining superhuman ILD sensitivity at low frequencies. These results suggest that phase-locked spike timing up to 1000 Hz is necessary for human-like dependence on binaural cues, implicating temporal coding in this aspect of perception.

Azimuth dependence of human localization requires phase locking

The non-human-like cue dependence under degraded phase locking was also evident in the dependence of localization acuity on azimuth. Human sound localization is best near the midline and becomes less accurate toward the periphery[66–68], as can be quantified by minimum audible angle thresholds[58] (the smallest detectable angular distance between two sources) (Fig. 3g). We simulated an experiment measuring minimum audible angle thresholds for pure tones. Thresholds measured from the 3000 and 1000 Hz phase locking models resembled those of human listeners (Fig. 3h). By contrast, the 320 and 50 Hz phase locking models exhibited a qualitatively different dependence on azimuth, with much higher thresholds away from the midline. These results suggest that ITD cues conveyed by precise spike timing are particularly important for accurate localization away from the midline. This idea is consistent with findings that ILDs are less reliable at lateral azimuths by virtue of varying nonmonotonically with azimuth[69], which might make ITDs critical for lateral localization.

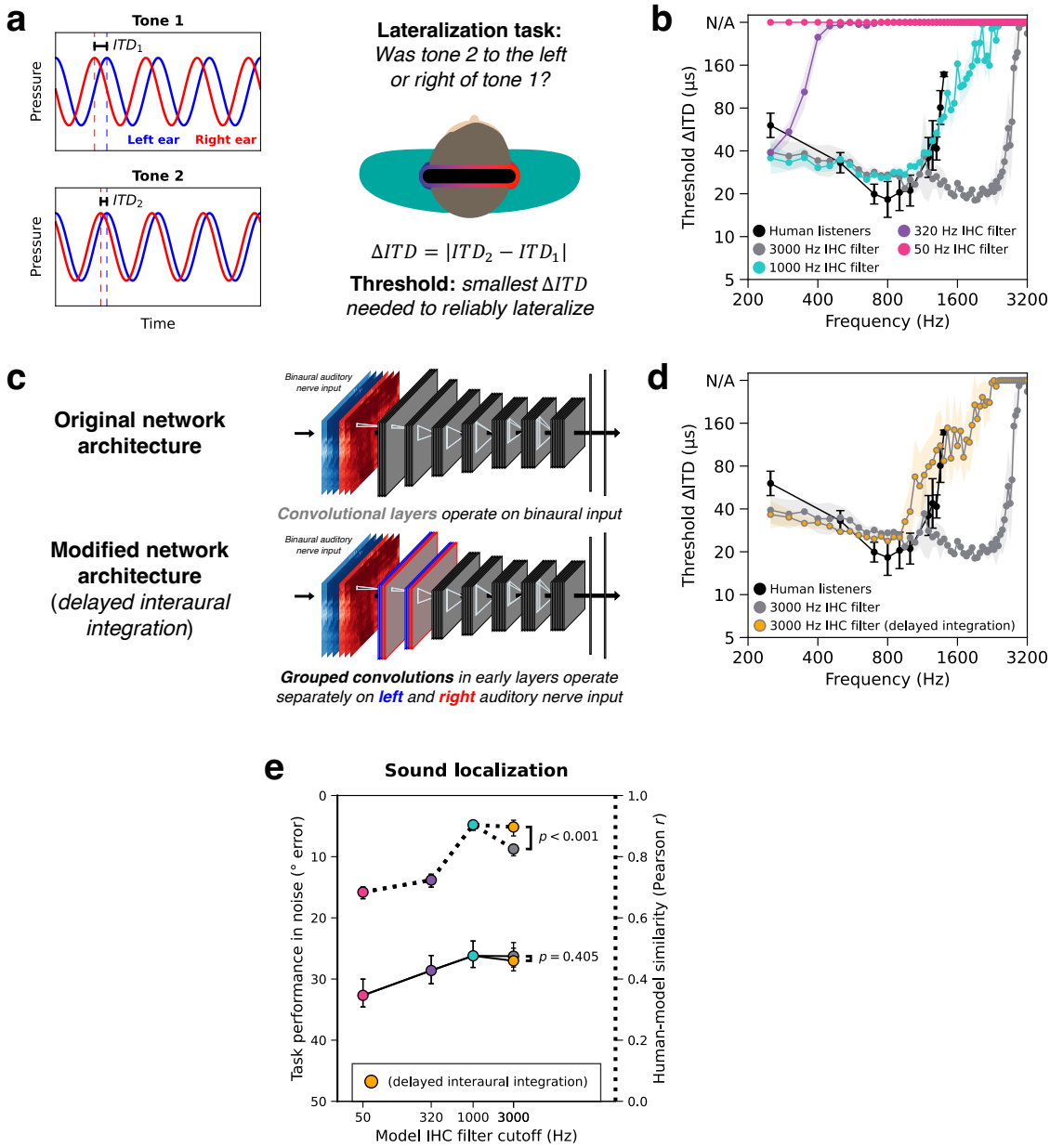


Fig. 4 | Upper frequency limit of interaural time difference sensitivity. a. Schematic of experiment used to measure ITD sensitivity as a function of frequency. On each trial, listeners heard a pair of pure tones with two different ITDs and judged whether the second tone was located to the right or left of the first. **b.** ITD lateralization thresholds measured as a function of frequency from humans and models. **c.** Schematic of neural network architecture modification to delay binaural integration. Replacing the first two convolutional layers with grouped convolutions (1 group for each ear) forces models to process the ears separately (and to downsample in time, reducing the fidelity of temporal coding, analogous to the loss of fidelity that occurs at each synapse in the auditory system) before binaural integration occurs in the first standard convolutional layer. Blue and red represent information from the left and right ears, respectively. **d.** ITD lateralization thresholds measured as a function of frequency from humans and models with and without the modified network architectures (both models had the same 3000 Hz phase locking limit in their auditory nerve representation). Error bars in b and d indicate ±2 standard errors of the mean across human participants or network architectures. Human data are re-plotted from the original study[45]. **e.** Effect of phase locking limit on sound localization in noise (left y-axis, solid lines) and human-model behavioral similarity (right y-axis, dotted lines). The data is re-plotted from Fig. 2a but now includes the delayed interaural integration model. The statistical significance of differences between models with and without delayed interaural integration was assessed by two-tailed paired comparisons. Error bars indicate 95% confidence intervals bootstrapped across network architectures.

Physiological model architecture constraints improve predictions of ITD sensitivity

Human listeners are remarkably sensitive to ITDs at low frequencies, but this sensitivity deteriorates at higher frequencies[44]. In principle this sensitivity could be limited by the upper limit of phase locking in the auditory nerve. However, human sensitivity instead declines rapidly above 1 kHz and is fully lost by 1.5 kHz[45] – well below the presumptive 3-5 kHz phase locking limit of the auditory nerve. To better understand this discrepancy, we studied the frequency limits of ITD sensitivity in our models.

ITD sensitivity as a function of frequency has been characterized with ΔITD thresholds with pure tones (single frequencies; Fig. 4a). In such experiments, listeners judge which of two lateralized tones (each with a different ITD) appears further to the right. The ΔITD threshold is the smallest change in ITD needed to reliably discriminate tones in this way. We simulated one such previously published experiment[45] and measured model ΔITD thresholds as a function of frequency. Model ΔITD thresholds were unmeasurably high for frequencies above a model's phase locking limit, as expected (Fig. 4b). Thresholds measured from the 1000 Hz phase locking network produced the closest match to human behavior. The 3000 Hz phase locking model in fact exhibited superhuman ITD sensitivity, with thresholds on the order of 20μs even up to 2.5 kHz. This discrepancy with humans suggests that some additional factor limits human ITD sensitivity.

Because ITD estimation requires a comparison of input from the two ears, perceptual sensitivity to high-frequency ITDs requires temporal coding at that frequency to be maintained in the auditory system until the stage at which this comparison is made[70,71]. One explanation for the lower limit of ITD sensitivity in humans is that anatomical constraints force information from each ear to pass through additional synapses before being compared, with some loss of temporal precision at each synapse. But this explanation in turn raises the question of why the auditory system would not have evolved a way to make the comparison happen earlier.

Because the models developed here can be tested in naturalistic conditions, they provide an answer to this question. When tested on naturalistic auditory scenes (natural sounds in noise), the 1000 Hz phase locking model localizes just as well as the 3000 Hz model (Fig. 2a and 3d). And across all other psychoacoustic experiments we simulated, there was no significant difference in human-model similarity between the 1000 and 3000 Hz phase locking models (Fig. 2b and Supplementary Fig. 2). These results suggest that temporal coding above ~1000 Hz provides little adaptive benefit, such that delaying the comparison incurs little cost.

To test the idea that "early" interaural comparisons accounted for our model's superhuman ITD sensitivity, we altered the neural network architectures slightly to delay interaural integration (Fig. 4c). We replaced the standard convolution operations in the earliest neural network stages with grouped convolution operations (with one group for each ear), such that the resulting models must initially process information from the left and right auditory nerve separately. Reasoning that synapses introduce temporal jitter that effectively imposes low-pass filtering[72], interaural integration was only allowed to occur in the models after early temporal pooling layers (see Methods). We trained these modified neural network architectures with 3000 Hz phase locking auditory nerve input and evaluated them on the full set of sound localization experiments. Consistent with our hypothesis, the models lost sensitivity to high-frequency ITDs (like humans; Fig. 4d) but were otherwise unaffected (Supplementary Fig. 2). Delaying interaural integration thus increased the overall human-model similarity score (which aggregates results across all experiments) for the 3000 Hz model ($p < 0.001,\ d = 12.0$, evaluated by bootstrapping across 10 neural network architectures; Fig. 4e, dotted line). These results indicate that additional physiological constraints can in some cases produce better matches to human behavior. Moreover, delaying interaural integration did not impair localization performance in noise ($p = 0.405$, $d = 0.783$, comparing localization error between the delayed and non-delayed 3000 Hz models; Fig.

4e, solid line). This latter result provides a normative explanation for the solution that biological auditory systems have arrived at over evolution, as it suggests there is little cost to real-world behavior when integration is delayed.

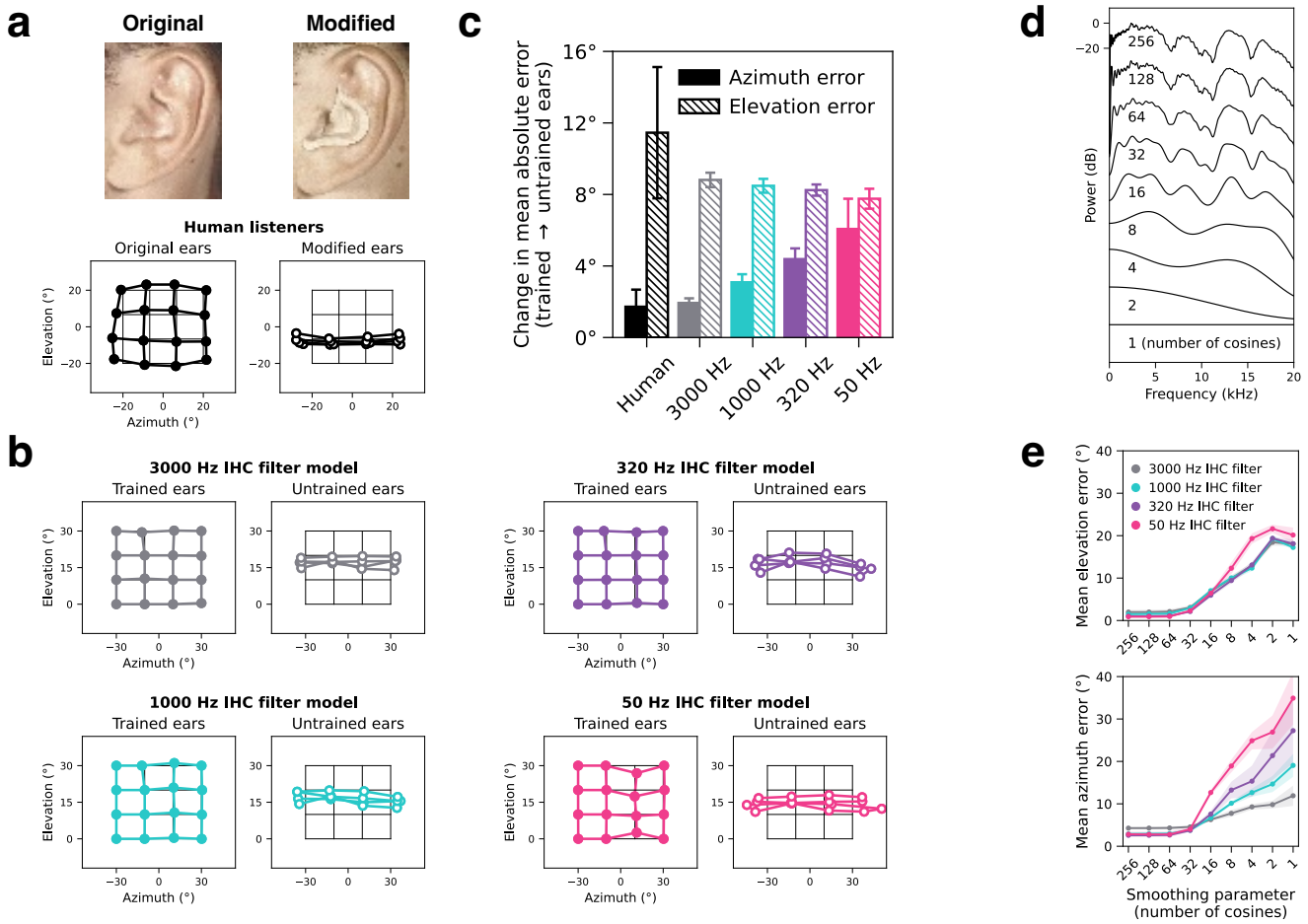


Fig. 5 | Localization models with degraded auditory nerve spike timing rely on spectral cues to judge azimuth as well as elevation. a. Photographs and human data from the original study manipulating pinna cues[57]. Hofman et al. (1998) measured human sound localization before and after inserting plastic molds into participants' ears to change the pinnae's direction-specific filtering. Sound localization judgments (thick lines, circle markers) with the participants original (left) and modified (right) ears are plotted as a function of azimuth and elevation, superimposed on a grid of the true locations (thin lines, no markers). Data were scanned from original results figures, and then aggregated across participants. **b.** Model results for an analogous experiment in which models were evaluated on sounds rendered with either the HRTFs used for training (trained ears) or a different set of HRTFs (untrained ears). Same conventions as in panel a. **c.** The increase in mean absolute azimuth and elevation error due to ear alteration for humans and for each model. Error bars indicate ±2 standard errors of the mean across human participants or network architectures. **d.** Spectral smoothing of pinna cues[56]. Power spectra of HRTFs progressively smoothed by lowering the number of cosines used to approximate the discrete cosine transform. **e.** Effect of the spectral smoothing manipulation on model accuracy in elevation and azimuth. Mean absolute elevation (top) and azimuth (bottom) errors are plotted as a function of the HRTF smoothing parameter used to render stimuli for the models. Error bars indicate ±2 standard errors of the mean across network architectures.

Models without access to phase locking exhibit inhuman spectral cue dependence

Whereas localization in azimuth is dominated by binaural cues, localization in elevation is mediated in large part by spectral cues from the pinnae of the outer ear (Fig. 5a). Manipulating these spectral cues (either physically by altering pinna shape with an ear mold[57] or virtually by altering the head-related transfer function used to spatialize a sound[56] via earphones) impairs elevation judgments in humans. These same manipulations have minimal effects on azimuth judgments.

We simulated these experiments on models optimized with different phase locking cutoffs. Our models were always trained using a set of head-related transfer functions (HRTFs) measured from a standard model of the human head and torso[73] (KEMAR). We evaluated the models with sounds rendered with altered HRTFs, thus simulating a change to the pinnae. We tested each model on 45 alternative sets of ears (HRTFs measured from the ears of 45 different people, as well spectrally smoothed versions of the KEMAR HRTFs used to train models, to test the spectral resolution of the cue dependence.

A priori it was not clear what to expect. The spectral cues used for elevation judgments are believed to reside at frequencies above the presumptive upper limit of phase locking[74], and thus their fidelity should not itself depend on phase locking. However, it seemed plausible that in the absence of phase locking an observer might become abnormally dependent on spectral cues to make up for impaired binaural information, and that this might indirectly provide another diagnostic signature of phase locking dependence.

When tested with alternative pinnae, elevation judgments collapsed in all models, as in human listeners with ear molds, indicating that spectral cues were used irrespective of phase locking (Fig. 5b). However, the effect of alternative ears on azimuth was different for different phase locking cutoffs. In human listeners and in models with high-fidelity temporal coding, changing pinnae had little effect on azimuthal accuracy. But in models with degraded temporal coding, azimuthal localization accuracy was worse with alternative pinnae, indicating that the absence of phase locking rendered models dependent on pinna cues for azimuthal localization, unlike humans (Fig. 5c). These results suggest that human-like dependence on ear-specific cues (i.e., only for elevation) emerges only when models have access to phase-locked spike timing.

This inhuman dependence of azimuthal localization on monaural spectral cues was also evident in the effects of removing spectral details from the cues (Fig. 5d). As the peaks and valleys of the trained HRTFs were parametrically smoothed away, model elevation judgments progressively collapsed, regardless of phase locking limit, as expected (Fig. 5e). By contrast, azimuth judgments were significantly more impaired by the smoothing in models with lower phase locking limits, suggesting they learned to use fine spectral details to localize in azimuth as well as elevation (unlike humans with normal hearing but consistent with the behavior of some single-sided cochlear implant users[75]).

Not all localization phenomena are inextricably linked to phase-locked spike timing

Although removing phase locking caused pronounced discrepancies with human behavior (Supplementary Fig. 2a-f), some behaviors were relatively unaffected. All models exhibited the "precedence effect", in which localization judgments are dominated by the initial part of a sound[76] (Supplementary Fig. 2g; evidently the models without phase locking learned to prioritize ILD cues from the onset of a sound in order to localize accurately amid reflections). All models also exhibited human-like dependences of localization accuracy on bandwidth[77] (Supplementary Fig. 2h) and of elevation accuracy on high-frequency spectral cues[74] (Supplementary Fig. 2i).

Model optimization -- word and voice recognition

To model speech perception, we optimized models to recognize words and voices using the Word-Speaker-Noise dataset[78] (Fig. 6a). This dataset consists of ~6 million 2-second speech excerpts superimposed on real-world background noise. Training on this dataset has previously been shown to produce models that yield the best current predictions of auditory cortical responses[79]. Models were jointly optimized to classify stimuli according to the word that appeared in the middle of the excerpt (794-way word recognition task) and the talker that produced the utterance (433-way voice recognition task). These tasks were intended to capture some of the challenges of everyday speech and voice recognition (background noise, large numbers of classes, high degree of variability), subject to practical constraints of dataset generation and model optimization (see Methods). We also trained models on each task individually and found similar results (Supplementary Fig. 3 and 4). We present results from the joint-task model here.

Phase locking improves voice but not word recognition in real-world noise

We first measured human and model word recognition performance as a function of SNR in different types of background noise (Fig. 6b). Lowering the phase locking limit produced negligible deficits for model word recognition accuracy in any of four tested noise classes: recorded auditory scenes, speech babble, instrumental music, and stationary speech-shaped noise. All models performed comparably to humans tested on the same stimuli, suggesting fine-grained temporal information provides little benefit for word recognition in noise, at least when spatial cues are unavailable.

We next measured the same models' voice recognition performance on the same stimuli (Fig. 6c). At low SNRs, models with access to phase locking performed better than models without. Almost all the benefit from phase locking incurred below 320 Hz, suggesting phase locking up to but not above the F0 of most human speech improves voice recognition in noise.

To further search for naturalistic noise conditions in which phase-locked spike timing might contribute to word recognition, we measured human and model word recognition in each of 43 different real-world auditory textures[80] (Fig. 6d; Supplementary Fig. 5). At a fixed SNR of -3 dB, these different textures produced a wide range of human word recognition scores (25% to 80% correct; this variation was highly reliable, with Spearman-Brown corrected split-half reliability = 0.968). Models tested on the same stimuli produced similar word recognition scores as humans, accounting for ~95% of the explainable variance in the human data. However, this similarity again held regardless of the phase locking limit (Fig. 6e). A scatter plot comparing word recognition scores between the 50 and 3000 Hz phase locking model (Fig. 6f, left) shows a small benefit of phase locking for word recognition (mean benefit = +2.3%; standard deviation = 2.3%). This result extends previous findings that neural networks trained to recognize speech can replicate patterns of human speech intelligibility[17,81–83], and further underscores the minimal benefit of high-fidelity phase locking for word recognition in noise (for monaural conditions in which localization cues cannot aid performance).

An analogous scatter plot of model voice recognition scores measured with the same stimuli (Fig. 6f, right) shows a considerably larger benefit than for word recognition (mean benefit = +14.9%; standard deviation = 5.7%). Comparing absolute voice recognition performance to that of humans is practically challenging because listeners do not all recognize the same voices and overall accuracy depends strongly on listener familiarity with voices. Instead, we asked whether the qualitative characteristics of human voice recognition were shared by the models, and whether this depended on the phase locking limit.

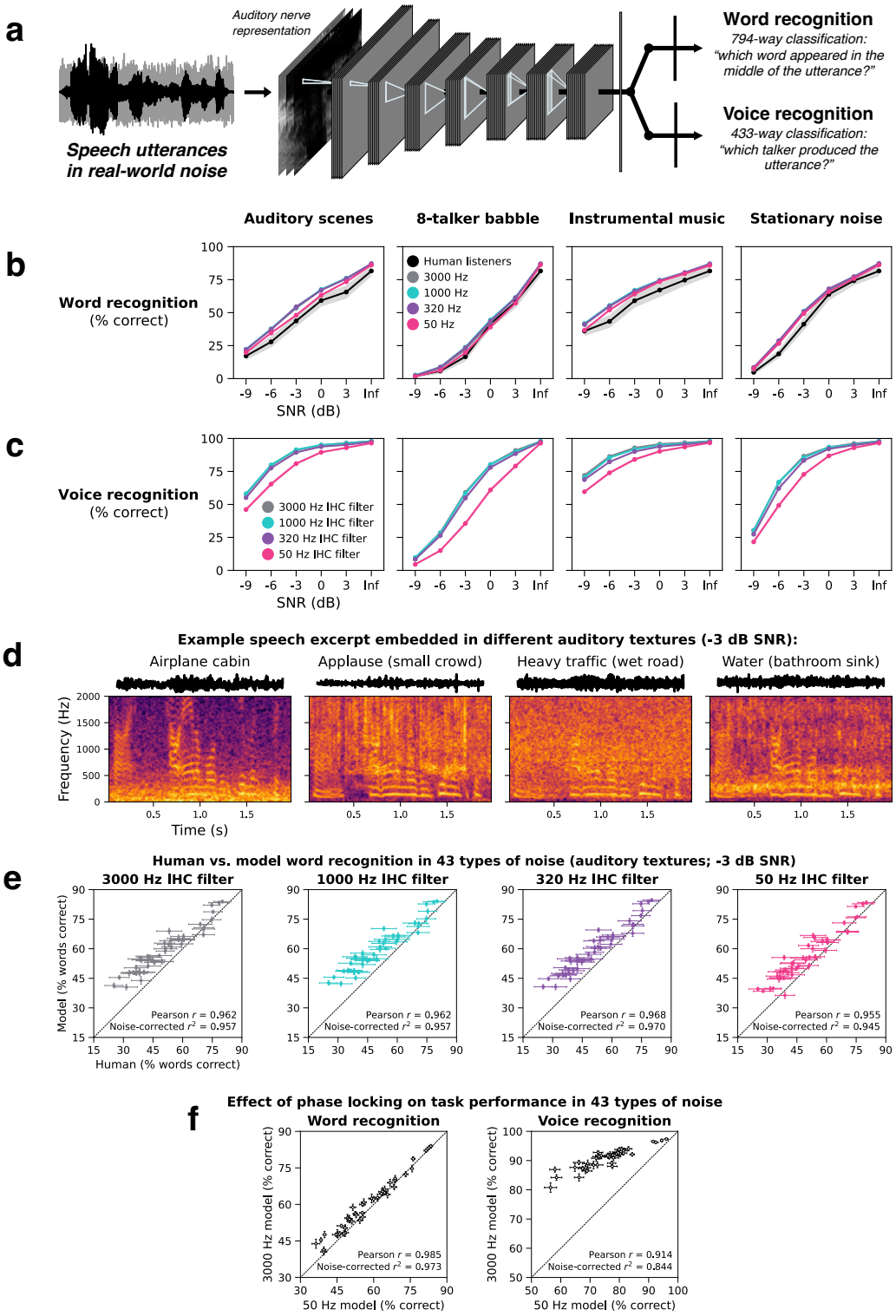


Fig. 6 | Auditory nerve spike timing improves voice but not word recognition in real-world noise. a. Speech model schematic. Deep artificial neural networks were jointly optimized to recognize words and voices from simulated auditory nerve representations of speech in noise. The two tasks shared all model stages up to the final task-specific output layers. **b.** Human and model word recognition as a function of SNR. Each panel plots task performance in a different naturalistic noise condition[17]. **c.** Model voice recognition as a function of SNR. **d.** Spectrograms of the same speech excerpt embedded in different auditory textures. **e.** Human vs. model word recognition scatter plots for speech embedded in each of 43 distinct auditory textures at -3 dB SNR. Each data point represents the human and model word recognition score for a single auditory texture. **f.** Effect of phase locking on model word and voice recognition in 43 distinct auditory textures. The left scatter plot compares word recognition performance for the 50 and 3000 Hz IHC filter models. The right scatter plot compares voice recognition performance for the 50 and 3000 Hz IHC filter models. All error bars indicate ±2 standard errors of the mean across human participants or network architectures.

The dependence of human voice recognition on absolute pitch requires phase locking

Humans rely in part on absolute pitch to recognize voices. When a familiar talker's voice is pitch-shifted or made inharmonic (by frequency-jittering its harmonic components to be inconsistent with any single F0) (Fig. 7a), the voice is less recognizable[84]. To assess whether these characteristics were shared by the models and if they depended on phase locking, we measured human and model voice recognition with pitch-shifted (Fig. 7b, closed symbols) and inharmonic speech (Fig. 7c, closed symbols). Models were tested on familiar voices (but held-out speech utterances) from the training set. Humans were tested on celebrity voices[84].

There is no way to match the relative familiarity of test voices between human and model participants, confounding comparisons of absolute performance. However, models with access to phase-locked spike timing best replicated the qualitative properties of human behavior. Human voice recognition was best for voices at their natural F0 and fell off with progressively larger shifts in either direction. Human performance was also impaired by making voices inharmonic. Models with the 50 Hz phase locking limit exhibited superhuman robustness to these pitch manipulations, suggesting the reliance on absolute pitch evident in humans only emerges with the aid of phase locking. The presumptive explanation is that pitch cues from phase locking aid performance in noise, such that models with phase locking learn a recognition strategy that uses these cues. This strategy leaves them dependent on these cues and thus produces worse performance when pitch is altered. By contrast, the model without phase locking does not use this cue and so is more robust to its alteration. These results provide additional evidence for a role of phase locking (up to ~320 Hz) in human voice recognition and pitch perception.

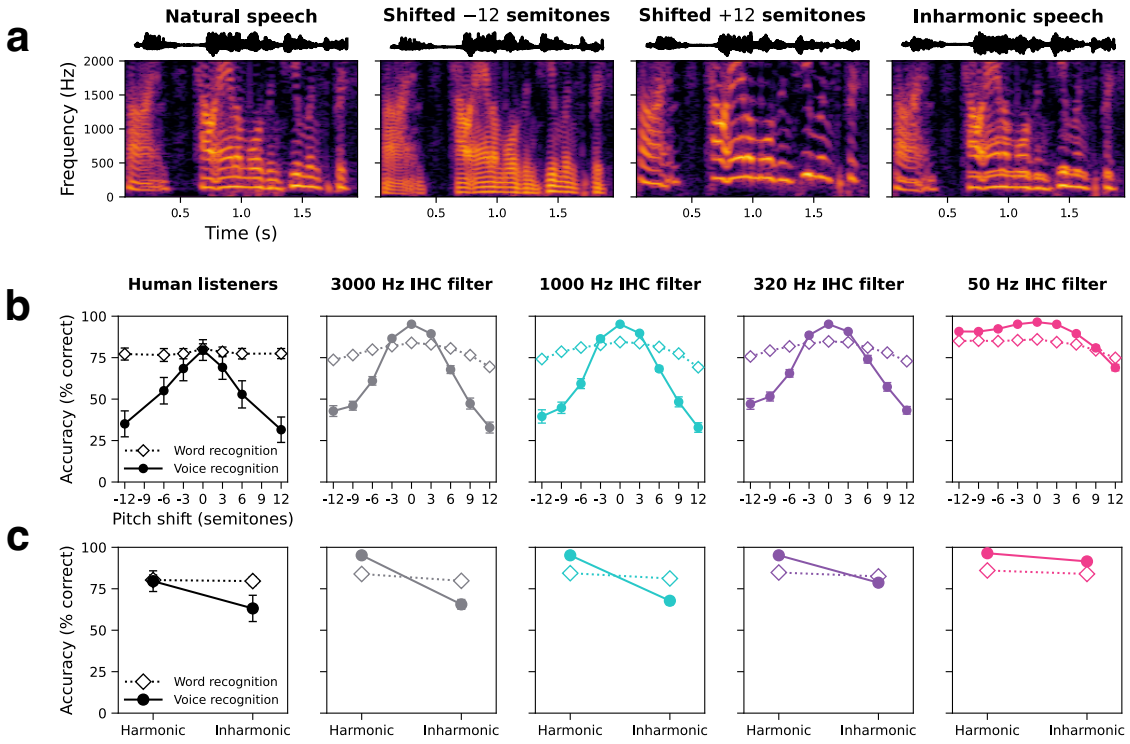


Fig. 7 | Auditory nerve spike timing is critical for human-like voice recognition. a. Stimuli for pitch-altered word and voice recognition experiments. Spectrograms show the same speech excerpt resynthesized in four different pitch conditions: unmodified (natural), pitch-shifted down 12 semitones, pitch-shifted up 12 semitones, and inharmonic. In the inharmonic condition, harmonic frequency components were randomly frequency-shifted such that they were no longer integer multiples of a common F0 and were no longer linearly spaced in frequency. **b.** Word and voice recognition accuracy for humans and models tested on pitch-shifted speech. **c.** Word and voice recognition accuracy for humans and models tested on harmonic and inharmonic speech. All error bars indicate ±2 standard errors of the mean across human participants or network architectures.

We also measured human and model word recognition with pitch-shifted and inharmonic speech (Fig. 7b and c, open symbols). In contrast to the results for voice recognition, human word recognition was unaffected by these pitch manipulations. Model performance was similarly robust regardless of phase locking limit, remaining comparable to humans in all conditions.

Phenomena previously linked to phase locking – effect of tone vocoding

In addition to general proposals that phase locking aids speech perception in noise, phase locking has been linked to two specific effects on human speech recognition. The first is the effect of tone vocoding – a signal manipulation intended to remove information conveyed by phase locking, which in some conditions produces deficits in speech intelligibility[39]. The second is the benefit of spatial separation between sound signals, which exhibits individual differences across human listeners, especially in reverberant listening conditions[40,85]. These individual differences have been proposed to be mediated by the integrity of temporal coding, presumably due to nerve fiber survival. We asked whether the models would exhibit these effects, and whether this depended on access to high-fidelity phase locking.

Tone vocoding is a stimulus manipulation in which a speech waveform is first decomposed into frequency bands with a cochlear filter bank[39] (Fig. 8a). The temporal envelopes of each band are then extracted and imposed on pure tone carriers at the center frequency of each channel and summed to produce a new waveform with similar envelope cues to the original but uninformative fine structure (because the tone carriers are constant over time and fixed across stimuli). Hopkins and Moore investigated the contribution of fine structure to speech intelligibility in noise by tone vocoding all frequency channels above a given cutoff. The authors increased this cutoff from 0 (all channels vocoded) to 32 (no channels vocoded), progressively increasing the upper frequency limit of fine structure information preserved in the stimulus. Speech reception thresholds were measured as a function of this cutoff in two types of noise (stationary and modulated; Fig. 8b). In stationary noise, thresholds improved somewhat as the cutoff increased. However, this improvement was more pronounced in modulated noise, with significant improvements up to 24 channels (corresponding to 4102 Hz; Fig. 8c, leftmost panel). This result was taken to suggest that humans benefit from the information in the monaural fine structure of sound waveforms, potentially upwards of 1000 Hz.

We tested our models on the same stimulus manipulation (Fig. 8c, right panels). Models with 3000 and 1000 Hz phase locking limits qualitatively and quantitatively replicated the human pattern of behavior, with speech reception thresholds improving as more high-frequency fine structure was preserved, particularly in modulated noise. The benefit from fine structure information was reduced in the 320 Hz phase locking model and fully eliminated by 50 Hz. This effect drove the lower overall human-model similarity for word recognition in the 50 Hz model (Fig. 2b).

These model results are consistent with the qualitative interpretation of the original human results[39] as implicating phase locking in this particular effect on speech intelligibility. However, they suggest that the frequency dependence of the tone vocoding manipulation is not directly related to the frequencies of phase locking used by the brain, contrary to the intention of the original manipulation. Specifically, word recognition in the models benefitted from added high frequency information beyond their respective phase locking limits. For instance, the 1000 Hz model received a benefit from frequencies well above 1000 Hz. Since the simulated auditory nerve representations cannot encode temporal fine structure so far above the phase locking limit, the task-relevant information contributed by high frequencies cannot be represented by high frequency phase-locked spike timing.

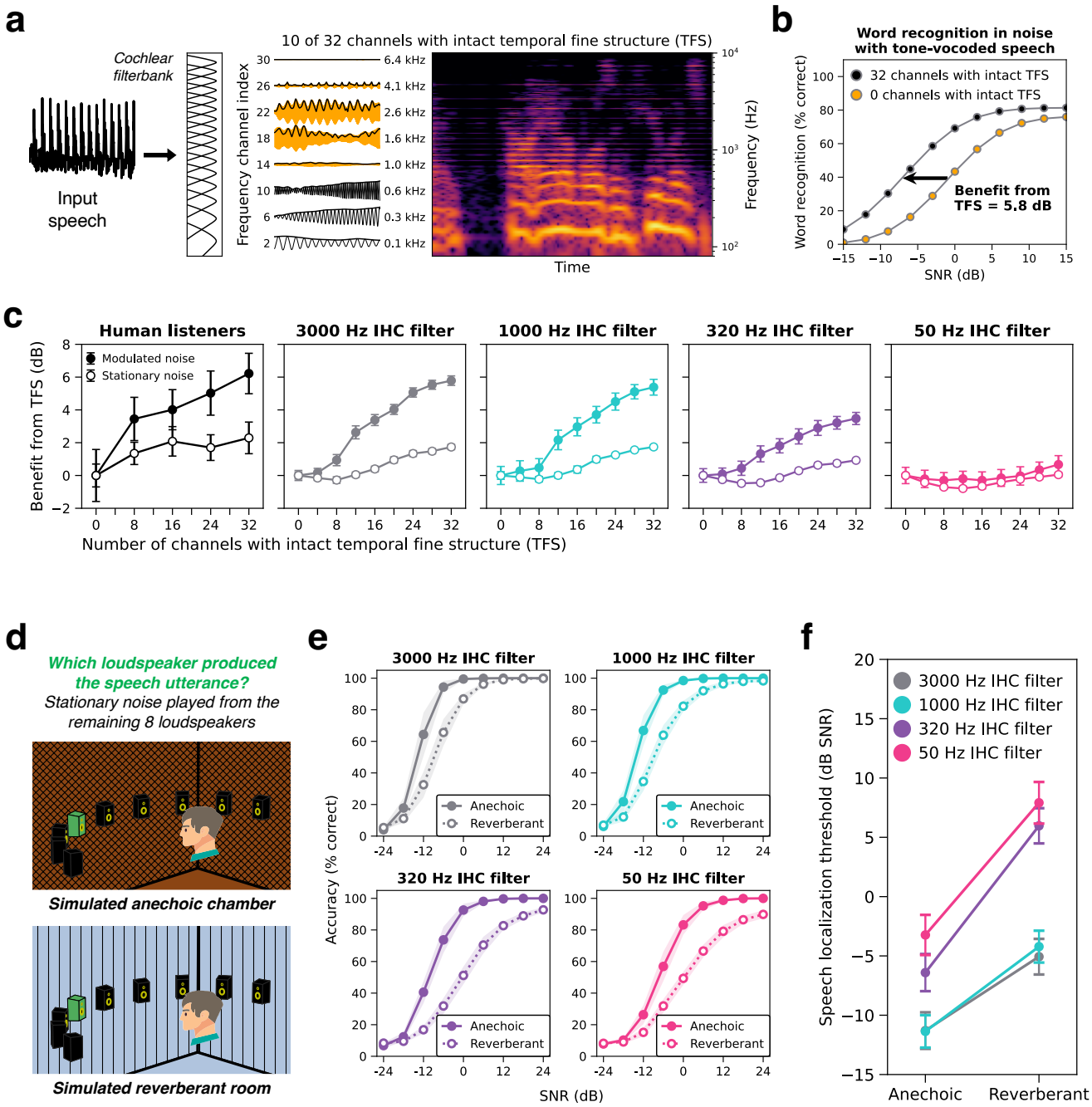


Fig. 8 | Auditory nerve phase locking is needed to account for phenomena previously linked to temporal fine structure. a. Schematic of tone-vocoding stimulus manipulation with a "cutoff channel" of 10. A speech waveform was separated into 32 frequency bands by a band-pass filter bank that mimics the cochlea's frequency tuning. Frequency channels up to and including the cutoff channel were left intact. In frequency channels above the cutoff, temporal fine structure (TFS) was disrupted by replacing the band with a pure tone carrier at the channel's center frequency, amplitude modulated by the envelope of the original band. **b.** The benefit from temporal fine structure was quantified by plotting word recognition accuracy vs. SNR and measuring leftward shifts in these psychometric functions as the cutoff channel (i.e., the number of channels with intact temporal fine structure) was increased. All shifts were computed relative to performance with fully tone-vocoded speech (0 channels intact, orange circles). **c.** Tone vocoding results. The benefit from temporal fine structure -- measured from humans and models -- is plotted as a function of the number of channels with intact temporal fine structure. Open circles plot the benefit in stationary noise and closed circles plot the benefit in amplitude-modulated noise. Human data in **c** is re-plotted from the original study[39] and errors bars indicate ±1 standard error of the mean across participants. **d** Schematic of the speech localization experiment in anechoic and reverberant conditions. **e.** Model sound localization accuracy as a function of SNR and reverberation. Panels plot performance in a simulated anechoic (solid symbols) and reverberant (open symbols) room for each phase locking model. **f.** The effect of phase locking and reverberation condition on speech localization thresholds measured from the psychometric functions in e. All model error bars indicate ±2 standard errors of the mean across network architectures.

One alternative explanation is that tone vocoding interferes with harmonic frequency relationships, such that when high-numbered harmonics are vocoded, they no longer produce temporal envelope variations at the F0 (which the auditory nerve encodes via phase locking to the F0). Because pitch is an important cue for sound segregation, disrupting the encoding of F0 could produce speech recognition deficits. Consistent with this alternative explanation, model word recognition exhibited very similar deficits in noise for inharmonic speech[86] as for tone-vocoded speech (Supplementary Fig. 6), with similar interactions with phase locking.

These results suggest phase-locked spike timing is needed to comprehensively account for human word recognition behavior. Yet as shown in Fig. 2a and 6, models optimized without fine timing information exhibited only minimally worse word recognition performance in naturalistic conditions (at least in the real-world noise conditions tested here). This combination of results suggests that the dependence of human-like word recognition on phase locking is a consequence of sharing machinery with tasks that benefit more from phase locking (perhaps voice recognition).

Phenomena previously linked to phase locking – localization of speech

Better encoding of temporal fine structure has also been proposed to correlate with the ability to direct spatial attention to voices in challenging acoustic environments[40,85]. Although our models did not possess selective attention, we could test whether phase locking would enable better localization of speech in noisy and reverberant environments, as would be necessary to direct spatial attention. We simulated a localization-in-noise experiment in which listeners report which of 9 loudspeakers (2m away, spanning -80° to +80° azimuth in 20° steps) produced a speech utterance, with threshold equalizing noise[59] played from the remaining 8 loudspeakers (Fig. 8d). We measured model performance as a function of signal-to-noise ratio in both a simulated anechoic chamber and in a moderately reverberant room (RT60 = 1s) (Fig. 8e). All models performed worse in reverberation, but the degraded 320 and 50 Hz phase locking models were particularly impaired, producing a significant interaction between phase locking and room condition ($F(3,36) = 75.27,\ p < 0.001,\ \eta^2_{partial} = 0.86$) (Fig. 8f). These results suggest that fine-grained temporal coding should help listeners attend to individual voices in challenging acoustic environments (e.g., the cocktail party problem), consistent with previous proposals[40,85]. The other major cue for selective attention in cocktail party scenarios is the sound of a target talker's voice, in particular the voice pitch[87,88]. Given that phase locking is needed for voice recognition and for the representation of the voice F0 (Fig. 7), our results collectively suggest that the cues for auditory attention are likely to be compromised without intact phase locking.

Replication with simplified cochlear model

State-of-the-art cochlear models that best capture the nonlinear response properties of the auditory nerve are computationally expensive, which can limit their integration into larger-scale models of the auditory system. To investigate whether the fine-grained details of these models are critical to account for naturalistic behavior, we also optimized models with a simplified cochlear front-end. The simplified front-end consisted of a linear cochlear filter bank followed by half-wave rectification and low-pass filtering (to impose an upper limit on phase locking), the output of which was passed through sigmoid functions approximating the rate-level functions of high-, medium-, and low-spontaneous-rate fibers[46]. These stages yielded a three-dimensional array of instantaneous auditory nerve firing rates with the same dimensions as the detailed auditory nerve model. The spike sampling procedure was identical for the simplified and detailed cochlear models. We repeated the temporal coding manipulation in the simplified cochlear model (by setting the low-pass filter cutoff to 3000, 1000, 320, and 50 Hz). In addition to testing the importance of a detailed cochlear model, the simplified model also served to rule out the

possibility that effects observed with the detailed cochlear model were driven by unintended nonlinear consequences of adjusting filter parameters rather than degraded temporal coding per se.

Models with the greatly simplified cochlear stage qualitatively and in most cases quantitatively replicated the results obtained with the highly detailed model of the auditory nerve (Supplementary Fig. 7, 8, and 9). This result suggests that the effects we saw with the detailed cochlear model are not due to unintended interactions between its components. The results also indicate that future work could use the simplified model in many settings without a cost.

DISCUSSION

We developed models of real-world sound localization, voice recognition, and word recognition by optimizing artificial neural networks to classify simulated auditory nerve representations of natural sounds. The resulting models closely replicated human behavior for natural sounds as well as for synthetic experimental stimulus manipulations, despite being optimized solely for task performance with natural sounds. To investigate the perceptual role of temporal coding in human hearing, we separately optimized models with lower auditory nerve phase locking limits. This manipulation had a larger effect on some tasks than others, impairing sound localization and voice recognition, while leaving naturalistic word recognition largely intact, even in noise. However, patterns of behavioral performance deviated from those in humans in at least some experimental conditions for each of the three task domains when the phase locking limit was too low (or too high, in one case). This combination of results provides evidence that phase locking is used in human perception, and suggests that both binaural and monaural mechanisms for extracting information from spike timing must exist in the auditory system, and that models of human hearing must operate on high temporal resolution input if they are to accurately capture behavior. But the results also provide a normative perspective on why temporal coding is used. In particular, any role in speech recognition seems likely to be driven by its importance for other tasks, such as voice recognition, where it provides clear performance benefits in real-world conditions, or by its importance for the two main attentional cues to speech (location and voice).

The comparison of machine learning models to humans also provided insight into additional biological constraints that influence human performance. Models whose neural networks had immediate access to the left and right auditory nerve representations exhibited superhuman sensitivity to interaural time differences (at much higher frequencies than is seen in humans). Simply altering the model architecture to require monaural processing stages (and thus some loss of temporal fidelity) before interaural integration produced more human-like behavior, accounting for results in every experiment we considered. In addition to yielding a more comprehensive model of human localization behavior, here the main contribution of our approach was to help understand why interaural comparisons have an upper frequency limit below the presumptive upper limit of phase locking in biological auditory systems. Our results suggest that the human auditory system did not evolve to use interaural time differences much above 1000 Hz because there is little benefit to sound localization in natural settings. This set of results illustrates how our modeling approach enables normative understanding of sensory physiology.

Relation to prior modeling work

Our approach draws inspiration from ideal observer theory, an early application of which was to investigate the role of temporal coding in hearing[2,12] in simpler settings. Siebert first derived optimal solutions to frequency discrimination given the information available in auditory nerve responses to synthetic tones. He showed that optimal observers using different features of neural coding yielded different patterns of performance, but the resulting models severely overestimated human performance.

Later instantiations of this approach applied to more realistic models of the nerve encountered similar issues[12]. The overestimation of human performance in such settings is perhaps unsurprising given that there is little reason to believe that the human auditory system has been optimized for discriminating pure tones. This is a central limitation of the classical ideal observer approach: the simple tasks and stimuli for which it is tractable to derive ideal observers are not those that likely drove biological optimization processes. And for the perceptual tasks that humans are plausibly optimized for (e.g., recognizing and localizing natural sounds)[89], the derivation of provably ideal observers is intractable. The present results show how the toolbox of contemporary deep learning provides an avenue to resurrect the ideal observer approach for real-world perceptual tasks. Even though the resulting models are almost surely not fully optimal for the tasks they are optimized for, they permit scientific inferences about the consequences of optimization under biological constraints.

Other previous models suggested neural mechanisms for the extraction of timing information[90–92]. The main limitation of mechanistic models is that they do not make extensive behavioral predictions, and thus are difficult to test in the absence of direct neural evidence for the mechanism in question. Our approach is complementary: we employ models that make behavioral predictions, enabling the role of timing information to be tested noninvasively (but at the expense of not addressing the underlying neural circuitry). Our results place some constraints on the underlying circuit mechanisms by revealing the range over which phase locking matters for real-world tasks. Phase-locked spike timing up to ~1000 Hz seems to be most critical for localization. Whereas for the word and voice tasks virtually all benefit of phase locking incurs between 50 and 320 Hz. These results implicate monaural mechanisms for extracting phase locking in the range of hundreds of Hz, consistent with at least one recent mechanistic proposal[93].

The models developed here extend recent efforts to apply deep learning to auditory modeling[17,19,20,94,54,79]. In addition to providing models with fairly realistic simulations of the peripheral auditory system (with an appropriate number of spiking nerve fibers), the present experiments demonstrate substantially more extensive evidence for close matches to human behavior than were available in previous work. In particular, we show that task-optimized models replicate human localization in noise (Fig. 3d), the upper frequency limit of ITD discrimination (Fig. 4b&d), patterns of word recognition performance across a large set of natural background noises (Fig. 6e), and patterns of voice recognition performance across pitch manipulations (Fig. 7b&c). These results bolster the evidence that much of auditory perception can be accounted for with task optimization.

Limitations

The most obvious limitation of our approach is that there is no guarantee that current deep optimization methods and model architectures converge on optimal task solutions. Does a model without access to phase locking fail to achieve human-like performance because precise spike-timing information is strictly necessary for the task or because the model is insufficiently optimized? In principle, more expressive model architectures and optimization methods could lead to more human-like models without precise spike-timing. We hedged against this possibility in two ways. First, we used multiple neural network architectures for each model class, ensuring that the reported results do not reflect the idiosyncrasies of any single network architecture. Second, our conclusions do not hinge solely on differences in absolute performance. We also compared the pattern of human and model behavior across a broad range of psychoacoustic experiments, allowing us to identify qualitative differences in how models solve tasks given different types of peripheral input. The best models achieved consistently good qualitative matches across a large set of experiments. Even though models with different phase locking limits similarly accounted for human performance for some tasks in natural conditions, only some of these models (typically those with higher phase locking limits) exhibited human-like cue dependencies, implicating phase locking in the perceptual strategies employed by humans.

Like ideal observers, optimized machine learning models can also outperform human observers. Some of this may be attributed to human attentional lapses during experiments. Biological systems also may have sources of noise that are absent in our models. Apart from spike sampling in the auditory nerve input, our models were deterministic. Ideal observers often posit decision stage noise to bring model performance down to the level of humans[2,12], and the same logic and approach could be applied to machine learning models in the future.

Our approach relies on optimizing models for the "right" constraints, which by hypothesis are the tasks that are critical in daily life. We optimized models for tasks that are plausibly important in this sense, but the specific instantiations of the tasks were constrained by practical considerations. For instance, the word recognition task is limited to identifying the middle word in a speech clip, with a vocabulary that is large in absolute terms (794 words) but still substantially smaller than the vocabulary of typical humans. Similarly, our sound localization task was restricted to static sources in simplified rooms, due to constraints of the head-related transfer functions and virtual acoustic simulator used. These tasks are more realistic than those used in previous generations of models, but it is possible that the demands of tasks that are even more realistic would alter the results. We also built separate models for sound localization and speech perception (again due to practical constraints). It is possible that the demands of having to perform multiple tasks with the same auditory system could affect the results. We addressed this concern to some extent by training one model on both word and voice recognition. We found results that were similar to those obtained from separate models optimized for each of the tasks individually, with distinct effects of phase locking on the two tasks evident in both cases. But particularly given proposals that the fidelity of interaural time differences relates to difficulties hearing in noise because of its effect on spatial attention[85,40], more complicated models that concurrently localize and recognize speech could provide additional insights.

Our approach is most diagnostic when there are differences in human-model similarity across model conditions, and this is a function of the experiments used to assess behavioral similarity. When human-model behavioral similarity for a task does not vary across phase locking limits, it is difficult to exclude the possibility that other behavioral assays might show a difference. Specifically, we found little difference in human-model similarity for voice and word recognition for the three highest phase locking cutoffs. There is thus no evidence in our present results that phase locking above 320 Hz is used for speech perception, but it remains possible that some other experiment could reveal a distinction (akin to that seen for sound localization). For instance, models optimized to estimate fundamental frequency require phase locking upwards of 1000 Hz to account for human pitch perception[19], raising the possibility that speech-related experiments that assess more fine-grained use of pitch could show effects of higher phase locking cutoffs. One way to address this in the future could be to use the models to derive stimulus conditions that produce different results for different phase locking cutoffs, and then to compare these results to those in humans.

Future directions

We have treated the neural network stages of our models as a black box. Our models thus offer insight into which neural cues underlie perception but do not reveal how these cues are extracted by biological neural circuits[90,95]. In principle one could probe the tuning of units within the neural network, or relate the internal representations of different model stages to those in the brain[17,96,97,79]. The absence of realistic neural components in the models presented here limits the relevance of such analyses to hypotheses for actual neural circuits for extracting temporal information. However, future models that combine machine learning with specific mechanistic proposals have exciting potential to make progress on these questions. For instance, proposals for how brainstem circuitry may extract task-relevant cues[11,98] could be built into a machine learning model as an additional stage that is either fully fixed, or

that has a small number of tunable parameters. Task-optimizing a deep network back-end in such a model and asking if the resulting model accounts for behavior could help test mechanistic hypotheses. The representations in such models could also be compared to brain representations[17,79], or to human EEG and ABR measures proposed as diagnostics of temporal processing[99,100].

Our approach has natural extensions for modeling sensorineural hearing loss and cochlear implants. The healthy auditory peripheral stage in our models could be altered to simulate hair cell loss[101], cochlear neuropathy[102], or electrical stimulation from a cochlear implant[103] to reveal their effects on auditory behavior. Optimizing models with different hearing loss etiologies could yield insights into the diverse behavioral outcomes of individuals with hearing loss. We found that similarly accurate predictions of human behavior were possible with a greatly simplified cochlear model stage (through which gradients can be backpropagated). This raises the possibility of directly optimizing front-end processors[53] (or even individual sounds[78,104]) for perceptual outcomes in the model, which could be useful for developing hearing aids, cochlear implants, and diagnostic behavioral tests.

Another natural extension would be to investigate species differences[105]. For instance, owls are known to use phase locking well above 1000 Hz for localization[106,107]. This difference with humans could plausibly be driven by the smaller interaural time differences that result from a smaller head, potentially coupled with differences in the sounds owls and humans must localize. Models trained with head-related transfer functions from an owl, and training data generated from owl-relevant sounds, could provide insight into the pressures that give rise to such species differences.

The approach employed here could be applied to understand other biological factors that plausibly influence behavior, for instance stochasticity. Although sometimes posited to limit human performance, we know little about the function of apparent stochasticity in neural responses. The incorporation of stochasticity within models of real-world behavior offers an avenue to better understand the role of noise in perception.

The general approach of investigating neural coding features with models optimized for ecological tasks is not limited to hearing. Similar analysis of tactile perception could, for instance, elucidate the perceptual role of high-fidelity temporal coding in touch[108]. More generally, the use of machine learning to reveal the consequences of optimization under constraints has widespread potential for understanding links between biology and behavior.

METHODS

Peripheral auditory model

The Bruce et al. (2018) auditory nerve model[46] served as the primary peripheral front-end to our artificial neural networks. This model (henceforth referred to as the "detailed" auditory nerve model) was chosen because it captures many of the complex response properties of auditory nerve fibers and has been extensively validated against electrophysiological data from nonhuman animals[109–111,101,112,113]. Stages of peripheral signal processing in the model include: a fixed middle-ear filter, a nonlinear cochlear filter bank to simulate level-dependent frequency tuning of the basilar membrane, inner and outer hair cell transduction functions, and a synaptic vesicle release/re-docking model of the synapse between inner hair cells and auditory nerve fibers. Although the model's responses have only been directly compared to recordings made in nonhuman animals, the cochlear filter bandwidths in the version used in this paper were inferred from human behavioral and otoacoustic measurements[114].

The output of the auditory nerve model was a three-dimensional array of instantaneous auditory nerve firing rates with shape [N frequency channels, T timesteps, S fiber types]. Due to computational constraints, we simulated instantaneous auditory nerve firing rates at N=50 points along the cochlear frequency axis. Auditory nerve fiber characteristic frequencies were spaced uniformly on an ERB-number scale[115] between 125 and 16000 Hz for the localization model and between 128 and 8000 Hz for the speech model. The speech model's upper characteristic frequency limit was lower because some of the training data was derived from corpora with audio sampling rates of 16000 Hz (with a corresponding Nyquist limit of 8000 Hz), reflecting the common view that speech perception is dominated by cues below 8000 Hz. The higher upper limit of the localization model reflected the established existence of localization cues above 8000 Hz[74] (making it critical to use a high audio sampling rate, and to represent high audio frequencies, in a model of localization). The use of 50 frequency channels primarily reflects computational constraints (CPU time for simulating peripheral representations, storage costs, and GPU memory during training). In previous work we found that increasing the number of frequency channels tenfold had little effect on model behavior[19].The instantaneous firing rates within each channel were downsampled from 100 kHz (the nerve model's default sampling rate to which all audio was upsampled) to 10 kHz. The localization model operated on 1s inputs (T=10000). The speech model operated on 2s inputs (T=20000). The high sampling rates of auditory nerve responses ensured the information in high-frequency phase locking up to 3000 Hz could be represented. At each characteristic frequency, we simulated responses of S=3 different auditory nerve fiber types to represent canonical high (70 spikes/s), medium (4 spikes/s) and low (0.1 spikes/s) spontaneous rate fibers[47]. Fibers with different spontaneous rates vary systematically in their thresholds and dynamic ranges. High-spontaneous-rate fibers have the lowest thresholds but smallest dynamic ranges such that their firing rates saturate at conversational speech levels.

This array of instantaneous firing rates was then converted to an array of binomially sampled spike counts representing the population response of 32000 individual auditory nerve fibers per ear. This spike sampling is an innovation over previous deep neural network models of audition. The number of spikes occurring at each time-frequency-fiber bin was sampled from a binomial distribution with p = firing rate / sampling rate and n determined by the relative numerosity of different fiber types (n = fraction of fibers * 32000 total fibers / N frequency channels). We used 60% high-, 25% medium-, and 15% low-spontaneous-rate fibers[47]. To reduce the computational cost of sampling from 1.5 million ($N \times F \times T$) independent binomial distributions per ear and stimulus, we employed a Gaussian approximation. Rather than directly sampling spike counts from $Binomial(n,\ p)$, we instead sampled from $Normal(np,\ np(1-p))$ and rounded samples to the nearest integer, yielding an approximate sample from the desired binomial distribution.

Phase locking manipulation

The phase locking manipulation was identical to that introduced in our previous work[19]. We modified the upper frequency limit of phase locking in the auditory nerve by adjusting the cutoff frequency of the IHC low-pass filter within the auditory nerve model. By default, the low-pass characteristics of the IHC membrane potential are modeled as a 7th order filter with a cutoff frequency of 3000 Hz. We set this cutoff to 3000, 1000, 320 and 50 Hz.

Simplified cochlear model

The Bruce et al. (2018) auditory nerve model[46] is computationally expensive to run, requiring peripheral representations to be precomputed and stored on disk rather than generated on-the-fly during neural network optimization. Simulated auditory nerve representations of the training datasets alone required 12 TB (localization task) and 26 TB (speech tasks) per phase locking condition. We repeated experiments with a simplified cochlear model hard-wired into the neural network's computation graph, eliminating the need to precompute the nerve representation (we note that it might eventually be possible to instead approximate detailed auditory nerve models with neural networks trained for this purpose[116]). The simplified front-end consisted of a finite-impulse-response approximation to a gammatone cochlear filter bank (with impulse responses truncated to 50ms) followed by half-wave rectification and low-pass filtering to impose the upper limit on phase locking. Simplified cochlear models operated on 50 kHz audio for the localization task and 20 kHz audio for the speech tasks (the audio training data, which were the same as for models with the detailed cochlear stage, were upsampled to these rates for numerical convenience; they made downsampling easier on the GPU). Low-pass filtering in the simplified cochlear model was performed by convolving the rectified subbands with an impulse response measured from the Bruce et al. (2018) model's IHC filter (truncated at 50 ms and then Hanning windowed). The finite-impulse-response approximation of the IHC filter ensured the frequency-dependence of phase locking in the simplified cochlear model closely matched that of the detailed model (Supplementary Fig. 7a). After low-pass filtering, cochlear representations were downsampled to 10 kHz and passed through pointwise sigmoid functions approximating the rate-level functions of high-, medium-, and low-spontaneous-rate fibers. These sigmoid functions ranged from 0 to 250 spikes/s over a dynamic range of 20, 40, or 80 dB for high-, medium-, and low-spontaneous-rate fibers with respective thresholds of 0, 12, and 28 dB SPL. These stages yielded an array of instantaneous auditory nerve firing rates with the same dimensions as the detailed auditory nerve model. The spike sampling procedure was identical between the simplified and detailed cochlear models. We repeated the temporal coding manipulation in the simplified cochlear model by setting the IHC low-pass filter cutoff to 3000, 1000, 320, and 50 Hz.

Artificial neural network architectures

Simulated auditory nerve representations were passed as input to deep convolutional neural networks, each consisting of a series of feedforward layers. These layers were hierarchically organized and instantiated one of several simple operations: convolution with a linear kernel, pointwise rectification, pooling, normalization, linear transformation, dropout regularization, and softmax classification.

For each task, we used 10 distinct neural network architectures previously identified in large-scale random searches over architectural hyperparameters (e.g., number of layers, units per layer, convolutional kernel size and shape, and pooling extent). The individual architectures for each task are summarized in Supplementary Tables 1 and 2. For the localization model, we used the top 10 best performing architectures from Francl and McDermott (2022). These architectures implement pooling and normalization via max pooling and batch normalization operations[20]. For the speech models, we took the best-performing architecture (modified to use Hanning-weighted average pooling[78] and layer

normalization operations) from Saddler and Francl et al. (2021)[53] as a starting point. We then performed a local architecture search by making 20 new architectures via single hyperparameter modifications from the starting point (e.g., adding/subtracting one model stage, or changing the convolutional kernel shape in one layer at a time). We used the 10 best-performing networks from this local architecture search for the speech model architectures.

Localization network architecture modification

Unlike the speech models, localization models operated on binaural input. To provide this binaural input to the models, we concatenated the auditory nerve representations from the simulated left and right ear along the last axis of the input (that was also defined by the three nerve fiber types). This resulted in a single array with shape [N frequency channels, T timesteps, 2S fiber types]. Standard convolutional model stages have three-dimensional kernels, which allow the optimizable filters to integrate information across the last feature axis (i.e., between the ears in this case). Our default neural network architectures imposed no restriction on where in the model processing hierarchy interaural cues could be extracted. To test the effect of delaying interaural integration as happens in biological auditory systems, we replaced standard convolution operations in the earliest model stages with "grouped" convolutions. Grouped convolutions split their input representation along the feature axis and use a separate convolutional kernel filter for each group[117]. Setting the number of groups to 2 in the first convolutional layer separated the input for the left and right ear. Successive convolutional stages with 2 groups maintain separate monaural processing streams. Interaural integration then occurs at the first convolutional stage where the number of groups is set to 1 (i.e., standard convolution). To delay interaural integration in our networks until after significant temporal pooling had occurred, we set the number of groups to 2 for all convolutional stages prior to the point at which the representation was downsampled by a factor of at least 4 relative to the nerve model stage output (from 10 kHz to no greater than 2.5 kHz). The convolutional stages replaced with grouped convolutions in the delayed interaural integration models are highlighted in Supplementary Table 1.

Model optimization -- overview

Artificial neural networks were optimized to perform real-world hearing tasks operationalized as classification tasks. The training datasets and individual tasks are described in the subsequent sections. In general, training stimuli were labelled with a class (one label per task) and neural network parameters were iteratively updated to minimize the softmax cross-entropy loss function via stochastic gradient descent (ADAM optimizer) with gradients computed via back-propagation. Localization models trained for 200,000 steps with a batch size of 32 and learning rate of 0.0001. Word and voice recognition models trained for 400,000 steps with a batch size of 32 and learning rate of 0.00001. The learning rates were arrived at empirically as those that worked well for the task and architectures. Classification performance on held-out validation sets was recorded after every 10000 training steps. The neural network parameters producing the highest validation set performance during the training routine were used for the trained model. The number of steps in each model's training routine was chosen to obtain a plateau in validation set performance under all phase locking conditions. Model training times varied by architecture, but each model could be trained in 96 hours on a single NVIDIA A100 GPU on the MIT OpenMind Computing Cluster. Localization models typically trained in under 48 hours.

Model optimization -- sound localization

We used the sound localization task of Francl and McDermott (2022)[20] in which models classified noisy 1s auditory scenes according to the azimuth and elevation of a target natural sound. The source location classes spanned 360° in azimuth (5° bin width) and 0 to 60° in elevation (10° bin width), yielding a total of 504 output classes (72 azimuth × 7 elevation classes). To ensure that the task was well-

defined, the training scenes always consisted of a single natural sound rendered at one target location superimposed with real-world noise textures diffusely localized at 3 to 12 different distractor locations. Target sounds were taken from the Glasgow Isolated Sound Events[118] (GISE-51) subset of Freesound Dataset 50k[119] (FSD50K), which consists of variable-length recordings of individual sources spanning 51 categories of everyday sounds. We only used source clips for which the original 44.1 kHz sampling rate audio could be found in FSD50K (to ensure that spectral localization cues could be rendered faithfully). Our training and validation datasets were generated from 12465 and 1716 unique source clips, respectively. This was a substantial increase in target sources compared to the previous modeling work of Francl and McDermott (2022)[20], with the goal of increasing the robustness of the resulting model. For model evaluation and human experiments, we used 460 sounds from the GISE-51 evaluation set equally distributed across 46 sound categories (discarding 5 categories with fewer than 10 evaluation clips).

Texture-like background noise was sourced from a subset of the Audioset[120] corpus screened to remove nonstationary sounds (e.g., speech or music). The screening procedure involved measuring auditory texture statistics[80] (envelope means, correlations, and modulation power in and across cochlear frequency channels) from all recordings, and discarding segments over which these statistics were not stable in time, as in previous studies[121,122]. The screening procedure yielded 26515 and 562 unique 10s noise clips for the training and validation datasets, respectively. Auditory scenes for the training and validation data were constructed by combining randomly sampled pairs of target sounds and texture-like noise samples (sliced into 1s segments). As in previous work[20], we augmented the number of unique target waveforms by applying a randomly generated band-pass filter to the target in 50% of training and validation examples. Band-pass filter center frequencies were sampled log-uniformly between 160 and 16000 Hz. Bandwidths were sampled log-uniformly between 2 and 4 octaves and the filter order was drawn uniformly between 1 and 4. Individual target and noise sources were first spatialized and then summed together at SNRs uniformly drawn between -15 and +25 dB, except for 5% of scenes which included no noise. For all localization experiments, SNR referred to the target sound's level relative to the sum of all background noise sources.

To spatialize scenes, we used a virtual acoustic room simulator[123] to render sets of binaural room impulses responses (BRIRs) for a KEMAR in 2000 unique listener environments. The simulator used the image-source method and incorporated KEMAR's HRTFs[73]. We randomly generated 2000 unique listener environments by sampling different shoebox rooms (varying in size and wall materials) and listener positions (x, y, z coordinates and head angle) within each room. Room lengths and widths were sampled log-uniformly between 3 and 30m and room heights were sampled log-uniformly between 2.2 and 10m. The listener's head position was sampled uniformly in each room, subject to the constraint that the head was at least 1.45m from every wall and no higher than 2m from the floor. For each listener environment, we rendered BRIRs at 1008 source locations (2 distances from the listener $\times$ 72 azimuths $\times$ 7 elevations). One of the distances was 1.4m for every BRIR. The other distance was independently sampled for each BRIR (drawn uniformly between 1m and 0.1m less than the distance from the listener to a wall). 1800 unique listener environments were included in the training set and the remaining 200 were used for validation. This was a substantial increase over the previous model by Francl and McDermott (2022)[20], whose data was generated from only 5 rooms. The final training and validation datasets consisted of 1,814,400 and 201,600 binaural auditory scenes, respectively. Target natural sounds were placed once at each of the 2000 $\times$ 1008 source locations to ensure the dataset was balanced across the 504 target location classes. Auditory scenes were presented to the model during training at sound levels drawn uniformly between 30 and 90 dB SPL.

Model optimization -- word and voice recognition

The same dataset was used for the word and voice recognition training tasks. We used an augmented version of the Word-Speaker-Noise dataset[78], which consists of 230,356 unique speech excerpts embedded in 718,625 unique nonspeech background noise excerpts from Audioset[120]. Randomly sampled pairs of 2s speech and noise excerpts were combined to yield a training dataset of 5.8 million examples. A validation set of 370,000 examples was similarly constructed from speech and noise excerpts excluded from training. Each example was labeled with the talker that produced the speech utterance and the word that appeared in the middle of the utterance (i.e., that overlapped the 1s mark of the 2s utterance). The datasets contained 433 unique talker labels and 794 unique word labels.

We chose this closed-set word recognition task in order to facilitate supervised learning with a human-relevant task. We assume that words constitute one of the output representations of human speech recognition, and so are a good choice of model output representation given the desire to optimize for biologically relevant tasks. However, due to Zipf's law, it is difficult to obtain large numbers of examples of infrequently occurring words. As a result, if one includes most English words it is challenging to generate training sets that are balanced across word labels (with similar numbers of examples of each class, as is advantageous for supervised learning). The main alternative at present would be to build a model using methods from contemporary machine speech recognition, which typically involves training systems to map audio to characters (with subsequent stages to derive word labels from character strings). Given that character strings seem a poor candidate for the output representation of human speech recognition, we opted to instead use a word recognition task with a vocabulary size for which we could assemble a balanced training set.

Training models for voice recognition is complicated by the fact that large speech corpora are often crowd-sourced online, with individuals contributing recordings of themselves reading passages or responding to prompts. Models optimized for talker classification using such corpora may pick up on non-voice cues that predict these labels (e.g., characteristics of the recording device or environment). To help ensure that models learned robust voice representations, we applied a set of randomly sampled audio manipulations to the speech to approximate the variable conditions in which human listeners encounter the same voice. In 25% of the dataset, speech excerpts were augmented to increase natural voice variability by applying small pitch (±0.5 semitones) and tempo shifts (±20%) or simulating whispering (less than 0.5% of examples) via the STRAIGHT algorithm[124]. In an independently drawn 25%, we applied commonly encountered audio distortions like band-pass / equalization filters, lossy audio compression / transmission, and dynamic range companding. In another independently drawn 5%, we replaced background noise with speech babble (between 12 and 36 talkers, uniformly sampled) to give the model some exposure to multi-talker situations. SNRs for the augmented speech excerpts were drawn uniformly between -10 and +10 dB and sound levels were drawn uniformly between 30 and 90 dB SPL.

We jointly optimized individual models to recognize both voices and words using this augmented dataset. Dual-task optimization was accomplished with a separate output layer for each task. All preceding network stages were shared between the two tasks and parameters were updated to minimize the sum of the softmax cross entropy loss from both tasks.

Model evaluation -- overview

For each task, we first evaluated model behavior in naturalistic conditions. Wherever possible, we tested humans on the same naturalistic tasks using the same stimuli. For each task, we then tested models on stimulus manipulations from the psychoacoustics literature and compared behavior to human results. Human-model behavioral similarity was quantified for each experiment and model by measuring Pearson correlation coefficients and root-mean-squared error between analogous human

and model data points. In each experiment, we present the results averaged across the 10 model architectures.

Localization model evaluation -- sound localization in noise

Human experiment. We measured the ability of human listeners to localize natural sounds in noise using a 19-by-5 array of loudspeakers arranged on a hemisphere with 2m radius. The array spanned the 180° in azimuth (frontal hemifield) and 0° to 40° in elevation (10° spacing in both azimuth and elevation) relative to the listener's head at the center. 11 normal-hearing listeners (5 female) with ages between 21 and 30 each performed 460 trials with 460 unique target natural sounds from the GISE-51 evaluation dataset. On each trial a target natural sound was played from one of the 95 loudspeakers while threshold equalizing noise[59] played concurrently from 9 distinct loudspeakers. Target and noise locations were randomly sampled on each trial. The listener's task was to report which loudspeaker produced the target by entering the loudspeaker's label on a keypad. Listeners were instructed to direct their head at the speaker directly in front of them for the duration of the stimulus. Once the stimulus ended, they could look at the speaker where they thought the target had occurred to obtain the label. Listeners then redirected their head toward the front speaker prior to the start of the next trial. Experimenter observation indicated that participants were highly compliant with these instructions. Target sounds were presented at 60 dBA and noise levels were determined such that the SNR of the target relative to the sum of the 9 noise sources was -13.6, -6.8, 0, +6.8, or +13.6 dB. All stimuli were sampled at 44.1kHz and were 1s in duration, including 15ms onset and offset ramps (Hanning window).

Model experiment. Models were tested on all combinations of the 460 target natural sounds, 5 SNRs, and 95 target locations (218,500 total stimuli) used in the human experiments. Sources were spatialized in a virtual rendering of the speaker array room human listeners were evaluated in. To match the task between human and models, we restricted model localization judgments to azimuth and elevations corresponding to the 95 loudspeaker locations.

Human-model comparison. Human and model performance was quantified by measuring mean absolute spherical error (great circle distance), azimuth error, and elevation between the true and reported target sound location (plotted in Fig. 3d). Human-model similarity scores report the correlation (or root-mean-squared difference) between these human and model error metrics as a function of SNR.

Localization model evaluation -- psychoacoustics

We simulated an expanded version of the battery of localization experiments used in Francl and McDermott (2022)[20]. 6 of the 8 psychoacoustic experiments we used were identical to those in Francl and McDermott, using the same stimuli and analyses. The minimum audible angle and ITD lateralization experiments were the two additions, included because they seemed potentially relevant to the use of phase locking. All psychoacoustic stimuli for model localization experiments were sampled at 44.1 kHz.

ITD / ILD cue weighting

Human experiment. We simulated the experiment of Macpherson and Middlebrooks (2002), which measured shifts in perceived azimuth for virtual sounds with additional ITDs and ILDs imposed[55]. In the original experiment, 13 participants (5 female) were played sounds over headphones and reported perceived azimuth by turning their head to face the virtual source. The experiment took place in an anechoic chamber and used both low-pass (0.5 to 2 kHz) and high-pass (4 to 16 kHz) 100ms noise bursts with 1ms squared-cosine ramps at the onset and offset.

Model experiment. We used identical stimuli spatialized in a virtual anechoic room at 0° elevation and 0° to 360° azimuth in steps of 5°. For each of the source locations and noise bands, we also separately created ITD- and ILD-biased versions of the stimuli. ITD-biased versions were generated by imposing additional ±300μs and ±600μs time delays between the two ears. ILD-biased versions analogously imposed additional ±10 and ±20 dB level differences between the two ears. We collected model azimuth predictions for each stimulus. Azimuth predictions in the rear hemifield were mapped to the frontal hemifield by reflecting across the coronal plane. We compared the model azimuth prediction for each ITD- and ILD-biased stimulus ("the biased azimuth") to the azimuth prediction for the corresponding unbiased stimulus ("the unbiased azimuth"), computing shifts in the biased azimuth relative to the unbiased azimuth. Azimuth shifts for ITD-biased stimuli were expressed in μs by subtracting the ITD of a real source at the biased azimuth from the ITD of a real source at the unbiased azimuth. Azimuth shifts for ILD-based stimuli were expressed in dB by subtracting the ILD of a real source at the biased azimuth from the ILD of a real source at the unbiased azimuth. Expressing azimuth shifts in cue units enables calculation of a dimensionless perceptual weight by dividing the azimuth shift by the imposed cue amount. Separate ITD and ILD perceptual weights were computed for low-pass and high-pass noise by averaging across all azimuths and bias magnitudes. For instance, an ITD perceptual weight of 1 indicates that, for a given virtual stimulus, an additional ITD of τ shifts the perceived azimuth by an angle corresponding to an ITD change of τ between two real source locations. Perceptual weights of 0 indicates that imposing additional ITDs or ILDs has no effect on the perceived the azimuth. These weights are plotted for each frequency range and cue type in Fig. 3f.

Human-model comparison. Human-model similarity was quantified by comparing ITD and ILD perceptual weights measured for low-pass and high-pass noise between humans and models.

Minimum audible angle vs. azimuth

Human experiment. Mills (1958) measured human localization acuity as a function of frequency and azimuth by playing pure tones to a blindfolded listener from a rotating boom in an anechoic chamber. Minimum audible angle thresholds were defined as the smallest change in azimuth required for the listener to discriminate whether a tone's location shifted left or right between two presentations. The key result was that localization acuity was best near the midline and degraded steadily towards the periphery. Mills (1958) only reported thresholds from a single human listener[58], but the general result is well-established and holds across different experimental paradigms[68].

Model experiment. We measured model thresholds by simulating a left/right lateralization experiment. Pure tones (1s duration including 70ms half-Hanning windows at onset and offset) were spatialized in a virtual anechoic room at 0° elevation and azimuths of -90° to +90° in steps of 0.5° (using linear interpolations of BRIRs spaced 5° apart). For each tone, we collected the model's predicted location probability distributions. These distributions were then multiplied by a mask assigning zero probability to nonzero elevations and azimuths outside the frontal hemifield (intended to replicate a human participant's knowledge that the tones to be discriminated were at this subset of all possible locations). This resulted in probability distributions over predicted azimuth in the frontal hemifield for each stimulus. Left/right discrimination trials were simulated by comparing the means of these distributions for pairs of stimuli rendered at different azimuths. We used the mean of the distribution rather than the maximum because the mean provided a fine-grained location estimate that allowed for more precise discrimination thresholds. Trials in which the signed predicted azimuth of the second tone was larger than the signed predicted azimuth of the first were counted as rightward judgments. Minimum audible angle thresholds for different frequencies (250, 500, 750 and 1000 Hz) and reference azimuths (0° to 75° in steps of 5°) were inferred from psychometric functions (proportion of rightward shifts as a function of azimuth difference) constructed from all possible trials within ±10° azimuth of the reference for each frequency. Model minimum audible angle thresholds were defined as the azimuth difference that

yielded 70.7% rightward shifts (calculated by fitting Normal cumulative distribution functions to the psychometric functions).

Human-model comparison. We averaged human and model thresholds across pure tone frequencies of 250, 500, 750, and 1000 Hz to yield a single results plot of accuracy vs. azimuth (plotted in Fig. 3h). Human-model similarity was quantified by comparing average model thresholds with linearly interpolated human thresholds as a function of absolute azimuth between 0° and 75°.

ITD lateralization vs. frequency

Human experiment. The upper frequency limit of fine structure ITD sensitivity in humans has classically been measured by asking listeners to make left/right lateralization judgments with pure tones presented over headphones. The pure tones have identical envelopes (by using a fixed window to eliminate onset ITDs) and zero ILD (identical amplitude between the two ears). Listeners hear pairs of tones with different ITDs and judge whether the second tone sounded left or right of the first. The ΔITD threshold is the smallest change in ITD between two tones that a listener can reliably discriminate in this way. We simulated the experiment of Brughera et al. (2013), who measured ΔITD thresholds of 4 young adult listeners (1 female) with 250, 500, 700, 800, 900, 1000, 1200, 1250, 1300, 1350, and 1400 Hz pure tones[45].

Model experiment. We simulated the experiment on models by measuring predicted location probability distributions from models tested on 500ms pure tone stimuli (including 100ms linear onset and offset ramps). Fine structure ITDs ranged from -160μs to 160μs in steps of 1μs. Frequencies ranged from 50 to 4000 Hz in steps of 50 Hz. ΔITD thresholds were inferred from model judgments using the same method as the minimum audible angle experiment. Model predictions were compared for pairs of stimuli with different ITD, and rightward judgments were assigned to trials for which the signed azimuth prediction was larger for the second tone than the first. Psychometric functions were constructed for each frequency (proportion of rightward shifts as a function of ΔITD) and the ΔITD threshold was defined as the difference in azimuth yielding 70.7% rightward shifts. These thresholds are plotted vs. frequency in Fig. 4b and 4d.

Human-model comparison. Human-model similarity was quantified by comparing log-transformed model thresholds with linearly interpolated human thresholds as a function of frequency between 250 and 1500 Hz.

Effect of changing ears

Human experiment. We simulated a change in our models' ears intended to be analogous to the manipulation of Hofman et al. (1998)[57]. In the original experiment, 4 participants localized white noise bursts presented in a 4-by-4 grid uniformly tiling ±20° in azimuth and ±20° elevation. Participants reported perceived locations by making eye movements to the source. After collecting baseline azimuth and elevation judgments, plastic molds were inserted in the participants ears, which altered the direction-specific filtering of their pinnae. Participants then repeated the localization task with modified ears[57].

Model experiment. We simulated the experiment by collecting baseline model azimuth and elevation judgments with the same stimuli used in the original human experiment (500ms noise bursts with a frequency band of 0.2 to 2 kHz) and then switching out the KEMAR HRTFs the model was trained with for 45 different sets of HRTFs from the CIPIC dataset[125]. Model azimuth and elevation predictions were collected for stimuli spatialized on 4-by-4 grid uniformly tiling ±30° in azimuth and 0° to 30° in elevation. Azimuths and elevations were not matched exactly to the human experiments due to constraints of the

available HRTFs. We averaged model judgments across the 45 different sets of HRTFs not used to train the model to compare against human judgments with modified ears.

Human-model comparison. To summarize the effects of changing ears on azimuth and elevation accuracy, we computed changes in mean absolute azimuth and elevation error with the untrained ears relative to the trained ears, averaging across all 16 source locations, yielding the graph of Fig. 5c. Human-model similarity was quantified by comparing absolute azimuth and elevation errors as a function of grid position and ear condition (as shown in Fig. 5a and 5b) between humans and models.

Effect of smoothing spectral cues

Human experiment. We simulated a modified version of the experiment originally conducted by Kulkarni and Colburn (1998), which measured the effect of HRTF spectral details on sound localization[56]. In the original experiment, 4 listeners were played white noise bursts in an anechoic chamber. Sounds were presented from either a physical loudspeaker in the room or virtually over open-backed headphones. The virtual sounds were spatially rendered at the loudspeaker's location using the participant's own HRTFs. Participants were tasked with reporting whether the sound came from the loudspeaker or the headphones. When the participants' full HRTFs were used, performance was at chance (50%). As the spectral details of the HRTFs were smoothed out by approximating the HRTF's discrete cosine transform with progressively fewer cosines, performance rose above chance as participants no longer perceived the virtual stimuli at the loudspeaker's location.

Model experiment. We applied the same smoothing manipulation to the KEMAR HRTFs (by retaining only the first 256, 128, 64, 32, 16, 8, 2, or 1 coefficients of the HRTF's discrete cosine transform) and evaluated model performance in a virtual anechoic room using 1s broadband (0.2 to 20 kHz) noise bursts. Model localization judgments were collected for each smoothing condition at 413 locations spanning 0° to 60° in elevation and 0° to 360° in azimuth (spacing determined by the locations of the measured KEMAR HRTFs). We computed mean absolute azimuth, elevation, and spherical errors as a function of the number of cosines used to approximate the HRTFs (plotted in Fig. 5e and Supplementary Fig. 2f).

Human-model comparison. Reasoning that higher absolute localization errors in the model would correspond to better performance on the human real/virtual discrimination experiment, we quantified human-model similarity by measuring the Pearson correlation between model absolute spherical error and human percent correct scores as a function of the smoothing parameter (Supplementary Fig. 2f).

Precedence effect

Human experiment. We simulated an experiment originally conducted by Litovsky and Godar (2010), which measured localization accuracy for 25ms (including 2ms cosine onset and offset ramps) pink noise bursts played at two different locations[126]. The bursts were played from two loudspeakers in an array spanning -60° to +60° in azimuth (20° spacing, 0° elevation) and were delayed relative to one another by 5, 10, 25, 50, or 100ms. The lag click was always presented at 0° azimuth and the lead click was presented variably at one of six azimuths (±20°, ±40°, ±60°) on each trial. 10 listeners (all female) with ages between 19 and 26 were tasked with reporting whether they heard one or two sounds as well the loudspeaker that produced each sound. Root-mean-squared azimuth errors were calculated separately for the lead and lag noise burst and reported as a function of the delay between the lead and lag bursts.

Model experiment. We evaluated models on the same stimuli rendered in a virtual anechoic room. Our models always reported a single location which we used to compute the root-mean-squared azimuth

error relative to both the lead and lag burst. The "precedence effect" refers to several different phenomena that occur when two sounds are played in close succession from different locations[127]. The model judgments can reflect one of these (the localization dominance of the leading sound), but does not allow investigation of others (fusion), because the models cannot report the presence of two sound source locations. Human and model results are plotted in Supplementary Fig. 2g.

Human-model comparison. Human-model similarity was quantified by comparing human and model azimuth error for both the lead and lag burst as a function of the inter-burst delay.

Bandwidth dependency of localization

Human experiment. We simulated the experiment of Yost and Zhong (2014), which measured the effect of stimulus bandwidth on localization accuracy with an array of 8 loudspeakers positioned between -15° to +90° in azimuth (15° spacing) relative to the midline[77]. 33 participants (26 female) with ages between 18 and 36 were tasked with reporting which loudspeaker produced a 200ms (including 20ms squared cosine onset and offset ramps) sound. Stimuli were pure tones or band-pass filtered white noise bursts with bandwidths of 1/20, 1/10, 1/6, 1/3, 1, and 2 octaves. Pure tone and center frequencies were set to 250, 2000, and 4000 Hz. Human listeners made 20 localization judgments per bandwidth, center frequency, and loudspeaker position.

Model experiment. We evaluated our models on simulations of the stimuli from the original human experiment, rendering sounds in a virtual anechoic room at azimuths of -90° to +90° in steps of 5°. Model localization judgments were restricted to the frontal hemifield and 0° elevation. Human and model results are plotted in Supplementary Fig. 2h.

Human-model comparison. Human-model similarity was quantified by comparing human and model root-mean-squared error as a function of bandwidth (averaged across center frequencies).

Median plane spectral cues

Human experiment. We simulated a modified version of the experiment by Hebrank and Wright (1974), which measured the accuracy of human elevation judgments as a function of the frequency content of noise bursts[74]. In the original experiment, 10 participants were played 1s noise bursts from a vertical array of speakers along the median plane spanning -30° to +210° in elevation with 30° spacing (0° is frontal). The experiment took place in an anechoic chamber and participants were tasked with reporting which speaker produced the noise burst. Noise bursts were either low-pass or high-pass with varying cutoff frequencies: 3.9, 6.0, 8.0, 10.3, 12.0, 14.5 or 16.0 kHz for the low-pass noise and 3.8, 5.8, 7.5, 10.0, 13.2 or 15.3 kHz for the high-pass noise.

Model experiment. We evaluated our model on noise bursts with the same cutoff frequencies rendered in a virtual anechoic room at elevations of 0°, 30°, 60°, 120°, 150°, and 180° along the median plane. To match the task between human and models, we restricted model localization judgments to azimuth and elevations along the median plane (model localization judgments corresponded to the highest softmax probability location class with azimuth 0° or 180°). Human and model results are plotted in Supplementary Figure 2i.

Human-model comparison. Human-model similarity was quantified by comparing human and model percent correct scores as a function of noise type and frequency cutoff.

Speech model evaluation -- word and voice recognition as a function of SNR and noise type

Human word recognition experiment. We measured human word recognition accuracy as a function of SNR in four different types of background noise[17] using an evaluation set of 376 unique speech excerpts (held out from the model training and validation sets). The experiment was a replication of an experiment by Kell et al. (2018)[17], modified to use words in the vocabulary of the models presented in this paper. The four types of noise were "auditory scenes", speech babble, instrumental music, and speech-shaped noise. The experiment also included a fifth noise condition which was not analyzed here. For the first three conditions, background noise excerpts (376 per condition) were sourced from IEEE AASP CASA Challenge[128] (auditory scenes), CommonVoice[129] (8-talker speech babble), and MUSDB18[130] (instrumental music). The babble was generated by summing speech excerpts from 8 different talkers. Speech-shaped noise was synthesized for each evaluation speech clip by imposing the power spectrum of the same speech clip on white noise. Speech excerpts were combined with background noise from each condition at 6 SNRs (noiseless and -9, -6, -3, 0, +3 dB) yielding 9024 possible stimuli (376 speech excerpts × 4 noise type × 6 SNRs). Stimuli were 2s in duration and sampled at 20 kHz. Individual participants heard only 376 stimuli (each unique speech excerpt was presented once), uniformly sampled across SNRs and noise types. To match our model word recognition task, participants were asked to report which word (from a list of 793) appeared in the middle of the utterance (defined as overlapping the 1s mark). Participants typed responses into a textbox and, as they typed, the displayed list of 793 words was filtered to include only words that matched the entered string. Only responses from the word list could be submitted. The experiment was run online and included 44 participants (13 female, 30 male, 1 nonbinary) who self-reported normal hearing, passed a headphone check[131], completed at least 100 trials, and responded correctly to at least 85% of catch trials intended to make sure they were paying attention to the experiment (isolated words presented in silence). Participants ages were between 18 and 62 (median 36) years. We did not run an analogous voice recognition experiment as there was no way to test humans and models in the same way (because every human is familiar with an idiosyncratic set of voices).

Model word and voice recognition experiment. We measured model word and voice recognition accuracy as a function of SNR and noise type using the same stimuli as the human experiment. Each of the 376 speech excerpts had a word and voice label included in the training dataset (376 unique words from 164 unique talkers). For the model experiment, the speech level was held fixed at 60 dB SPL and noise levels were adjusted to the desired SNRs. Models were evaluated on the full evaluation set (9024 stimuli). Results are plotted in Fig. 6b and 6c.

Human-model comparison. We quantified human-model similarity by comparing human and model word recognition accuracy as a function of SNR and noise types.

Speech model evaluation -- word and voice recognition in naturalistic auditory textures

Human word recognition experiment. To probe human speech-in-noise at a larger scale (to obtain a stronger model comparison test), we measured human word recognition accuracy in 43 different naturalistic auditory textures. The 376 evaluation set speech excerpts were embedded in 376 unique exemplars of each auditory texture. The 2s texture exemplars were previously generated[20] to match the statistics of 43 recorded real-world textures and the success of the iterative synthesis algorithm[80] was determined both subjectively (synthesized exemplars sounded like the recorded textures) and objectively (mean-squared errors between synthetic and original texture statistics were at least 40 dB below the mean-squared texture statistics of the original recordings). The experiment was identical to the previous word recognition experiment, except excerpts were randomly assigned to one of the 43 texture conditions and the SNR was held fixed at -3 dB. The online experiment included 47 participants (24 female, 23 male) who self-reported normal hearing, passed a headphone check[131], completed at least 100 trials, and responded correctly to at least 85% of catch trials (isolated words presented in silence). Participant ages were between 23 and 59 (median 39) years.

Model word and voice recognition experiment. We measured model word and voice recognition accuracy in each of the 43 auditory textures using the same stimuli as the human experiment. Models were evaluated on the full evaluation set (16168 stimuli = 376 speech excerpts × 43 auditory textures). Human and model results are plotted in Fig. 6e and 6f.

Human-model comparison. We quantified human-model similarity by comparing human and model word recognition accuracy as a function of the 43 auditory textures. The noise-corrected explained variance was calculated by dividing the human-model Pearson r^2 by the geometric mean of the human and model split-half reliabilities (after Spearman-Brown correction).

Word and voice recognition with pitch-altered speech

Human voice recognition experiment. We ran a modified version of the voice recognition experiment of McPherson and McDermott (2018), which measured human listeners' ability to recognize pitch-altered voices from famous celebrities[84]. Stimuli were 4s speech excerpts from 37 recognizable politicians, actors, singers, and television hosts. In the first block of 37 trials, each participant heard all 37 voices randomly assigned to one of 8 pitch-manipulation conditions (inharmonic or shifted ±12, ±6, ±3, 0 semitones from the original F0). In the second block of 37 trials, each participant heard different excerpts of the same 37 voices with no pitch shift. Each participant's results were analyzed only for first the block, limited to just the celebrity voices they successfully recognized in the second block and identified as familiar in a pre-experiment survey (an attempt to only measure performance for familiar voices, to make for a fairer comparison with the models). All stimuli were resynthesized with the STRAIGHT algorithm[124]. Voices were made inharmonic by shifting harmonic frequency components above the fundamental by random amounts uniformly sampled between -50% and +50% of F0[84,86]. Jitter values were sampled independently for each harmonic frequency and voice clip but were constrained (via rejection sampling) such that adjacent harmonics were always separated by at least 30 Hz. The experiment was a 100-alternative forced-choice task. Participants entered responses into a textbox which filtered a displayed list of 100 celebrity names and descriptors (e.g., "Dolly Parton (country singer-songwriter)") until there was only a single match, which the participant could submit. This procedure deviated slightly from the original McPherson and McDermott experiment, which was open set, and required the experimenter to then score participant's text responses by hand. The procedure adopted here standardized responses while maintaining some of the benefits of an open set experiment (minimizing the chance that participants would artifactually boost their scores by using a process of elimination with the list of possible voice choices). The experiment included 112 participants (46 female, 65 male, 1 nonbinary) who self-reported normal hearing and passed a headphone check[131]. Participant ages were between 20 and 73 (median 39) years. Because analysis was limited to voices for which participants demonstrated familiarity and each voice could only be assigned to one pitch condition, the number of participants for each condition ranged from 87 to 95.

Human word recognition experiment. We measured human word recognition accuracy for the same 8 pitch conditions using the 376 model evaluation set speech excerpts. The experiment was identical to the word recognition in noise experiments, except excerpts were randomly assigned to one of 8 pitch conditions (synthesized with the same procedure used for the voice recognition experiment) and presented in quiet. The online experiment included 22 participants (8 female, 14 male) who self-reported normal hearing, passed a headphone check[131], completed at least 100 trials, and responded correctly to at least 85% of catch trials (isolated and unaltered words presented in silence). Participant ages were between 25 and 70 (median 38) years.

Model word and voice recognition experiment. We measured model word and voice recognition accuracy on the pitch-manipulated evaluation set used in the human word recognition experiment. We

collect model word and voice predictions for the 376 speech excerpts in each of 10 pitch conditions (inharmonic or shifted ±12, ±9, ±6, ±3, 0 semitones from the original F0). Human and model results are plotted in Fig. 7b and 7c.

Human-model comparison. Human-model similarity across pitch conditions was quantified with separate correlation coefficients (or root-mean-squared error) for word and voice recognition. We compared human and model word or voice recognition scores as a function of the 8 shared pitch conditions.

Word recognition with tone-vocoded speech

Human experiment. We simulated an experiment originally conducted in humans by Hopkins and Moore (2009), which measured speech reception thresholds in stationary and modulated noise using progressively tone-vocoded speech[39]. In the original experiment, speech stimuli were split into frequency subbands with a 32-channel cochlear band-pass filter bank with center frequencies equally spaced on an ERB-number scale[115] between 100 and 10000 Hz. Frequency channels above a set cutoff channel (which determined the "number of channels with intact TFS") were tone vocoded, intended to disrupt temporal fine structure. Channels were tone vocoded by imposing the temporal envelope (absolute value of the Hilbert transform) of the original speech subband on a pure tone carrier at the channel's center frequency. Tone-vocoded subbands were band-pass filtered using the corresponding filter from the cochlear filter bank and summed together with the remaining unmodified subbands. The resulting stimuli were presented to listeners in both stationary and modulated speech-shaped noise. Modulated noise was amplitude-modulated with an 8 Hz sinusoid on a decibel scale with a peak-to-valley ratio of 30 dB. Human speech reception thresholds were measured from 10 normal hearing participants using an adaptive procedure that tracked the SNR needed to achieve 50% of words correct. Hopkins and Moore (2009) reported speech reception thresholds with the cutoff channel set to 0, 8, 16, 24, and 32.

Model experiment. We applied the same stimulus manipulation to our 376 evaluation set speech excerpts and measured model word recognition accuracy in stationary and modulated noise at SNRs of -15 to +15 dB in increments of 3 dB. We used the speech-shaped noise from the word recognition experiment describe above (and shown in Fig. 6b). Amplitude-modulated noise was generated by applying the same 8 Hz sinusoidal envelope used in the human experiment to the speech-shaped noise. Speech reception thresholds were calculated for the model by fitting a sigmoid to the psychometric function (word recognition accuracy as a function of SNR) for each condition and selecting the SNR that yielded half-maximal performance. We measured model speech reception thresholds with the cutoff channel set between 0 (all channels tone vocoded) and 32 (all channels intact) in steps of 4. Because our models were trained with speech sampled at 20 kHz, the 32-channel Gammatone filter bank used to synthesize model stimuli had center frequencies equally spaced on an ERB-number scale between 80 and 8000 Hz rather than 100 to 10000 Hz.

Human-model comparison. As in the original analysis by Hopkins and Moore, human and model speech reception thresholds for both noise types were expressed relative to speech with all channels vocoded (i.e., subtracted from the threshold with cutoff channel set to 0). Human-model similarity was quantified by comparing this "benefit from TFS" as a function of SNR and noise type between humans and models.

Speech localization in noise and reverberation

The purpose of this experiment was to evaluate the models' performance in another setting in which phase locking has been proposed to influence speech recognition – that in which spatial attention is used to select a target talker amongst distractor talkers[40,85]. We currently lack models that can perform

attentional selection tasks, such that it was not possible to conduct a model version of the published human experiments in this domain. Instead, we measured the effect of phase locking on the localization of speech, based on the logic that impaired localization of speech would translate to impaired spatial attention in cocktail party settings. We specifically tested conditions with noise and reverberation that have been found to produce individual differences in behavior that might be related to individual differences in the integrity of temporal coding in the auditory periphery.

Model experiment. Localization models were evaluated on the 376 evaluation set speech excerpts (nerve representations truncated to 1s) spatialized in both a virtual anechoic room and a virtual reverberant room. The reverberant (RT60 = 1s) room was a rendering of our loudspeaker array room from the localization-in-noise experiment described above (and shown in Fig. 3c and 3d). For the anechoic room, we changed all simulated room materials to be perfectly absorptive. Each speech clip was assigned to one of 9 loudspeaker locations (spanning -80° to +80° azimuth in steps of 20°) 2m from the simulated listener. On each trial, threshold-equalizing noise[59] played from the remaining 8 loudspeaker locations. The speech level was held constant at 60 dB SPL, and the total noise level was set to produce 9 different SNRs uniformly tiling -24 to +24 dB. Model localization judgments (restricted to the 9 possible locations) were collected for each of the resulting 6768 stimuli (376 speech excerpts × 2 reverberation conditions × 9 SNRs). Separate psychometric functions were constructed for each reverberation conditions by calculating the proportion of correct trials as a function of SNR. Speech localization thresholds were defined as the SNR yielding 70.7% trials correct, linearly interpolated.

Statistics. The statistical significance of the interaction between phase locking and reverberation conditions was assessed with a permutation test analogous to a mixed model ANOVA. An F-statistic was computed from the speech localization thresholds, with phase locking cutoff as a between-subjects factor and reverberation was a within-subjects factor. The F-statistic was re-computed 10000 times with permuted reverberation labels to assemble the null distribution used to calculate a p-value for the actual F-statistic.

Aggregate measures of task performance in noise

To summarize model task performance in noise (Fig. 2), we averaged model results across noise conditions for each task. For sound localization, we averaged mean absolute spherical errors across the 5 SNR conditions in Fig. 3d (-13.6, -6.8, 0, +6.8, and +13.6 dB). For word and voice recognition, we averaged recognition accuracy across the 5 SNR (-9, -6, -3, 0, and +3 dB) and 4 noise type conditions in Fig. 6b and c. We calculated 95% confidence intervals for each task performance summary metric by bootstrapping the model mean (sampling 10 neural network architectures with replacement 1000 times); these confidence intervals are plotted as the error bars on task performance in Fig. 2. The statistical significance of the effects of model manipulations (degraded phase locking or delaying interaural integration) on overall task performance in noise was assessed by comparing mean task performance metrics against bootstrapped null distributions from the 3000 Hz phase locking model task performance metrics. Two-tailed p values were estimated from Gaussian fits to these null distributions (as the probability of obtaining a score more extreme than that obtained from each degraded phase locking model under the null distribution). Effect sizes are quantified by measuring differences in the mean (Fig. 2a-c) or Cohen's d (Fig. 4e) between bootstrapped distributions of human-model similarity scores from different phase locking conditions.

Aggregate measures of human-model similarity

Human-model behavioral similarity was quantified separately for each model and experiment, first with a Pearson correlation coefficient. In each case, we compared mean model behavior (averaged across the 10 neural network architectures) with mean human behavior (averaged across experiment

participants). We calculated 95% confidence intervals for each human-model similarity value by bootstrapping the model mean (sampling 10 neural network architectures with replacement 1000 times). The statistical significance of the effects of model manipulations (degraded phase locking or delaying interaural integration) on overall human-model similarity was assessed by comparing mean human-model similarity scores against bootstrapped null distributions from the 3000 Hz phase locking model human-model similarity score (same analysis as for task performance metrics). Two-tailed p values are reported, and effect sizes were quantified by measuring Cohen's d between bootstrapped distributions of human-model similarity scores from different phase locking conditions.

To ensure conclusions were robust to the choice of similarity metric, we repeated human-model comparisons by measuring root-mean-squared error between analogous human and model data points. The analyses were identical to those with Pearson correlations, except root-mean-squared errors were z-scored across model conditions before averaging across experiments to account for different units and scales.

Error bars

Except where otherwise noted in figure captions, error bars in results figures indicate ±2 standard errors of the mean across 10 neural network architectures (model results) or across experiment participants (human results). There are two exceptions. The first is in plots of aggregate measures of model task performance and human-model similarity (Fig. 2a-c, Fig, 4e, Supplementary Fig. 1a-c, and Supplementary Fig. 7b-g), where error bars indicate 95% confidence intervals bootstrapped across 10 neural network architectures. The second exception is Fig. 3h, where the human error bars indicate ±2 standard errors from 1 listener (the original experiment's only participant[58]) averaged across 4 different pure tone frequencies.

ACKNOWLEDGMENTS

We thank Bertrand Delgutte, Jim DiCarlo, Michale Fee, and members of the McDermott lab for their helpful feedback on the manuscript. This work was supported by National Institutes of Health grant number R01DC017970 to J.H.M.

REFERENCES

1. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*. xi, 455 (John Wiley, Oxford, England, 1966).
2. Siebert, W. M. Frequency discrimination in the auditory system: Place or periodicity mechanisms? *Proc. IEEE* **58**, 723–730 (1970).
3. Barlow, H. B. The efficiency of detecting changes of density in random dot patterns. *Vision Res.* **18**, 637–650 (1978).
4. Geisler, W. S. Contributions of ideal observer theory to vision research. *Vision Res.* **51**, 771–781 (2011).
5. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
6. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nat. Neurosci.* **5**, 598–604 (2002).
7. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
8. Burge, J. & Geisler, W. S. Optimal defocus estimation in individual natural images. *Proc. Natl. Acad. Sci.* **108**, 16849–16854 (2011).
9. Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).
10. Goldstein, J. L. An optimum processor theory for the central formation of the pitch of complex tones. *J. Acoust. Soc. Am.* **54**, 1496–1516 (1973).
11. Dau, T., Püschel, D. & Kohlrausch, A. A quantitative model of the '"effective"' signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.* **99**, 3615–3622 (1996).

12. Heinz, M. G., Colburn, H. S. & Carney, L. H. Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. *Neural Comput.* **13**, 2273–2316 (2001).

13. May, T., van de Par, S. & Kohlrausch, A. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans. Audio Speech Lang. Process.* **19**, 1–13 (2011).

14. Guest, D. R. & Oxenham, A. J. Human discrimination and modeling of high-frequency complex tones shed light on the neural codes for pitch. *PLOS Comput. Biol.* **18**, e1009889 (2022).

15. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* **111**, 8619–8624 (2014).

16. Jozwik, K. M., Kriegeskorte, N., Storrs, K. R. & Mur, M. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* **8**, (2017).

17. Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630-644.e16 (2018).

18. Rajalingham, R. *et al.* Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).

19. Saddler, M. R., Gonzalez, R. & McDermott, J. H. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nat. Commun.* **12**, 7278 (2021).

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

21. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).

22. Mainen, Z. F. & Sejnowski, T. J. Reliability of spike timing in neocortical neurons. *Science* **268**, 1503–1506 (1995).

23. Marion-Poll, F. & Tobin, T. R. Temporal coding of pheromone pulses and trains in Manduca sexta. *J. Comp. Physiol. A* **171**, 505–512 (1992).

24. Victor, J. D. & Purpura, K. P. Nature and precision of temporal coding in visual cortex: a metric-space analysis. *J. Neurophysiol.* **76**, 1310–1326 (1996).

25. Cariani, P. A. & Delgutte, B. Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *J. Neurophysiol.* **76**, 1698–1716 (1996).

26. Carleton, A., Accolla, R. & Simon, S. A. Coding in the mammalian gustatory system. *Trends Neurosci.* **33**, 326–334 (2010).

27. Mackevicius, E. L., Best, M. D., Saal, H. P. & Bensmaia, S. J. Millisecond Precision Spike Timing Shapes Tactile Perception. *J. Neurosci.* **32**, 15309–15317 (2012).

28. Rose, J. E., Brugge, J. F., Anderson, D. J. & Hind, J. E. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *J. Neurophysiol.* **30**, 769–793 (1967).

29. Johnson, D. H. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J. Acoust. Soc. Am.* **68**, 1115–1122 (1980).

30. Palmer, A. R. & Russell, I. J. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hear. Res.* **24**, 1–15 (1986).

31. Cedolin, L. & Delgutte, B. Pitch of complex tones: rate-place and interspike interval representations in the auditory nerve. *J. Neurophysiol.* **94**, 347–362 (2005).

32. de Cheveigné, A. & Pressnitzer, D. The case of the missing delay lines: Synthetic delays obtained by cross-channel phase interaction. *J. Acoust. Soc. Am.* **119**, 3908–3918 (2006).

33. Verschooten, E. *et al.* The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints. *Hear. Res.* **377**, 109–121 (2019).

34. Rouiller, E., de Ribaupierre, Y. & de Ribaupierre, F. Phase-locked responses to low frequency tones in the medial geniculate body. *Hear. Res.* **1**, 213–226 (1979).

35. Liu, L.-F., Palmer, A. R. & Wallace, M. N. Phase-locked responses to pure tones in the inferior colliculus. *J. Neurophysiol.* **95**, 1926–1935 (2006).

36. Swaminathan, J. & Heinz, M. G. Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. *J. Neurosci.* **32**, 1747–1756 (2012).

37. Qin, M. K. & Oxenham, A. J. Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J. Acoust. Soc. Am.* **114**, 446–454 (2003).

38. Lorenzi, C., Gilbert, G., Carn, H., Garnier, S. & Moore, B. C. J. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc. Natl. Acad. Sci.* **103**, 18866–18869 (2006).

39. Hopkins, K. & Moore, B. C. J. The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *J. Acoust. Soc. Am.* **125**, 442–446 (2009).

40. Ruggles, D., Bharadwaj, H. & Shinn-Cunningham, B. G. Why middle-aged listeners have trouble hearing in everyday settings. *Curr. Biol.* **22**, 1417–1422 (2012).

41. Viswanathan, V., Shinn-Cunningham, B. G. & Heinz, M. G. Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble. *J. Acoust. Soc. Am.* **150**, 2664–2676 (2021).

42. Budak, M. *et al.* Contrasting mechanisms for hidden hearing loss: Synaptopathy vs myelin defects. *PLOS Comput. Biol.* **17**, e1008499 (2021).

43. Klumpp, R. G. & Eady, H. R. Some measurements of interaural time difference thresholds. *J. Acoust. Soc. Am.* **28**, 859–860 (1956).

44. Zwislocki, J. & Feldman, R. S. Just noticeable differences in dichotic phase. *J. Acoust. Soc. Am.* **28**, 860–864 (1956).

45. Brughera, A., Dunai, L. & Hartmann, W. M. Human interaural time difference thresholds for sine tones: The high-frequency limit. *J. Acoust. Soc. Am.* **133**, 2839–2855 (2013).

46. Bruce, I. C., Erfani, Y. & Zilany, M. S. A. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. *Hear. Res.* **360**, 40–54 (2018).

47. Liberman, M. C. Auditory-nerve response from cats raised in a low-noise chamber. *J. Acoust. Soc. Am.* **63**, 442–455 (1978).

48. Weiss, T. F. & Rose, C. Stages of degradation of timing information in the cochlea: A comparison of hair-cell and nerve-fiber responses in the alligator lizard. *Hear. Res.* **33**, 167–174 (1988).

49. Javel, E. & Mott, J. B. Physiological and psychophysical correlates of temporal processes in hearing. *Hear. Res.* **34**, 275–294 (1988).

50. Joris, P. X. & Verschooten, E. On the limit of neural phase locking to fine structure in humans. in *Basic Aspects of Hearing* (eds. Moore, B. C. J., Patterson, R. D., Winter, I. M., Carlyon, R. P. & Gockel, H. E.) 101–108 (Springer, New York, NY, 2013). doi:10.1007/978-1-4614-1590-9_12.

51. Verschooten, E., Desloovere, C. & Joris, P. X. High-resolution frequency tuning but not temporal coding in the human cochlea. *PLOS Biol.* **16**, e2005164 (2018).

52. Jacoby, N. *et al.* Universal and non-universal features of musical pitch perception revealed by singing. *Curr. Biol.* **29**, 3229-3243.e12 (2019).

53. Saddler, M. R. *et al.* Speech denoising with auditory models. in *Interspeech 2021* 2681–2685 (ISCA, 2021). doi:10.21437/Interspeech.2021-1973.

54. Feather, J., Leclerc, G., Mądry, A. & McDermott, J. H. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nat. Neurosci.* **26**, 2017–2034 (2023).

55. Macpherson, E. A. & Middlebrooks, J. C. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *J. Acoust. Soc. Am.* **111**, 2219–2236 (2002).

56. Kulkarni, A. & Colburn, H. S. Role of spectral detail in sound-source localization. *Nature* **396**, 747–749 (1998).

57. Hofman, P. M., Van Riswick, J. G. A. & Van Opstal, A. J. Relearning sound localization with new ears. *Nat. Neurosci.* **1**, 417–421 (1998).

58. Mills, A. W. On the minimum audible angle. *J. Acoust. Soc. Am.* **30**, 237–246 (1958).

59. Moore, B. C. J., Huss, M., Vickers, D. A., Glasberg, B. R. & Alcántara, J. I. A test for the diagnosis of dead regions in the cochlea. *Br. J. Audiol.* **34**, 205–224 (2000).

60. Batteau, D. W. & Huxley, H. E. The role of the pinna in human localization. *Proc. R. Soc. Lond. B Biol. Sci.* **168**, 158–180 (1967).

61. Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization*. (MIT Press, 1997).

62. Rayleigh, Lord. On our perception of sound direction. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **13**, 214–232 (1907).

63. Henning, G. B. Lateralization of low-frequency transients. *Hear. Res.* **9**, 153–172 (1983).

64. Hafter, E. R., Dye, R. H., Jr. & Gilkey, R. H. Lateralization of tonal signals which have neither onsets nor offsets. *J. Acoust. Soc. Am.* **65**, 471–477 (1979).

65. Klein-Hennig, M., Dietz, M., Hohmann, V. & Ewert, S. D. The influence of different segments of the ongoing envelope on sensitivity to interaural time delays. *J. Acoust. Soc. Am.* **129**, 3856–3872 (2011).

66. Makous, J. C. & Middlebrooks, J. C. Two-dimensional sound localization by human listeners. *J. Acoust. Soc. Am.* **87**, 2188–2200 (1990).

67. Carlile, S., Leong, P. & Hyams, S. The nature and distribution of errors in sound localization by human listeners. *Hear. Res.* **114**, 179–196 (1997).

68. Wood, K. C. & Bizley, J. K. Relative sound localisation abilities in human listeners. *J. Acoust. Soc. Am.* **138**, 674–686 (2015).

69. Macaulay, E. J., Hartmann, W. M. & Rakerd, B. The acoustical bright spot and mislocalization of tones by human listeners. *J. Acoust. Soc. Am.* **127**, 1440–1449 (2010).

70. Jeffress, L. A. A place theory of sound localization. *J. Comp. Physiol. Psychol.* **41**, 35–39 (1948).

71. Colburn, H. S. & Durlach, N. I. Models of binaural interaction. *Handb. Percept.* **4**, 467–518 (1978).

72. Grothe, B. & Sanes, D. H. Bilateral inhibition by glycinergic afferents in the medial superior olive. *J. Neurophysiol.* **69**, 1192–1196 (1993).

73. Gardner, W. G. & Martin, K. D. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am.* **97**, 3907–3908 (1995).

74. Hebrank, J. & Wright, D. Spectral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.* **56**, 1829–1834 (1974).

75. Grantham, D. W., Ricketts, T. A., Ashmead, D. H., Labadie, R. F. & Haynes, D. S. Localization by postlingually deafened adults fitted with a single cochlear implant. *The Laryngoscope* **118**, 145–151 (2008).

76. Wallach, H., Newman, E. B. & Rosenzweig, M. R. A precedence effect in sound localization. *J. Acoust. Soc. Am.* **21**, 468 (1949).

77. Yost, W. A. & Zhong, X. Sound source localization identification accuracy: bandwidth dependencies. *J. Acoust. Soc. Am.* **136**, 2737–2746 (2014).

78. Feather, J., Durango, A., Gonzalez, R. & McDermott, J. Metamers of neural networks reveal divergence from human perceptual systems. in *Advances in Neural Information Processing Systems* vol. 32 10078–10089 (2019).

79. Tuckute, G., Feather, J., Boebinger, D. & McDermott, J. H. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biol.* **21**, e3002366 (2023).

80. McDermott, J. H. & Simoncelli, E. P. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron* **71**, 926–940 (2011).

81. Spille, C., Ewert, S. D., Kollmeier, B. & Meyer, B. T. Predicting speech intelligibility with deep neural networks. *Comput. Speech Lang.* **48**, 51–66 (2018).

82. Weerts, L., Rosen, S., Clopath, C. & Goodman, D. F. M. The psychometrics of automatic speech recognition. Preprint at https://doi.org/10.1101/2021.04.19.440438 (2021).

83. Adolfi, F., Bowers, J. S. & Poeppel, D. Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Netw.* **162**, 199–211 (2023).

84. McPherson, M. J. & McDermott, J. H. Diversity in pitch perception revealed by task dependence. *Nat. Hum. Behav.* **2**, 52–66 (2018).

85. Ruggles, D., Bharadwaj, H. & Shinn-Cunningham, B. G. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proc. Natl. Acad. Sci.* **108**, 15516–15521 (2011).

86. Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H. & McDermott, J. H. Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nat. Commun.* **9**, 2122 (2018).

87. Bird, J. *et al.* Effects of a difference in fundamental frequency in separating two sentences. *Psychophys. Physiol. Adv. Hear.* 263–269 (1998).

88. Woods, K. J. P. & McDermott, J. H. Attentive tracking of sound sources. *Curr. Biol.* **25**, 2238–2246 (2015).

89. Kell, A. J. E. & McDermott, J. H. Deep neural network models of sensory systems: windows onto the role of task constraints. *Curr. Opin. Neurobiol.* **55**, 121–132 (2019).

90. Licklider, J. C. R. A duplex theory of pitch perception. *Experientia* **7**, 128–134 (1951).

91. Loeb, G. E., White, M. W. & Merzenich, M. M. Spatial cross-correlation: A proposed mechanism for acoustic pitch perception. *Biol. Cybern.* **47**, 149–163 (1983).

92. Shamma, S. & Klein, D. The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *J. Acoust. Soc. Am.* **107**, 2631–2644 (2000).

93. Joris, P. X. Entracking as a brain stem code for pitch: the butte hypothesis. in *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* (eds. van Dijk, P. et al.) 347–354 (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-25474-6_36.

94. Brochier, T. *et al.* From microphone to phoneme: an end-to-end computational neural model for predicting speech perception with cochlear implants. *IEEE Trans. Biomed. Eng.* **69**, 3300–3312 (2022).

95. Joris, P. X., Schreiner, C. E. & Rees, A. Neural processing of amplitude-modulated sounds. *Physiol. Rev.* **84**, 541–577 (2004).

96. Khatami, F. & Escabí, M. A. Spiking network optimized for word recognition in noise predicts auditory system hierarchy. *PLOS Comput. Biol.* **16**, e1007558 (2020).

97. Giordano, B. L., Esposito, M., Valente, G. & Formisano, E. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nat. Neurosci.* **26**, 664–672 (2023).

98. Carney, L. H. Supra-threshold hearing and fluctuation profiles: implications for sensorineural and hidden hearing loss. *J. Assoc. Res. Otolaryngol.* **19**, 331–352 (2018).

99. Bharadwaj, H., Verhulst, S., Shaheen, L., Liberman, M. C. & Shinn-Cunningham, B. Cochlear neuropathy and the coding of supra-threshold sound. *Front. Syst. Neurosci.* **8**, (2014).

100. Dai, L., Best, V. & Shinn-Cunningham, B. G. Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention. *Proc. Natl. Acad. Sci.* **115**, E3286–E3295 (2018).

101. Zilany, M. S. A. & Bruce, I. C. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J. Acoust. Soc. Am.* **120**, 1446–1466 (2006).

102. Furman, A. C., Kujawa, S. G. & Liberman, M. C. Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. *J. Neurophysiol.* **110**, 577–586 (2013).

103. Tabibi, S., Boulet, J., Dillier, N. & Bruce, I. C. Phenomenological model of auditory nerve population responses to cochlear implant stimulation. *J. Neurosci. Methods* **358**, 109212 (2021).

104. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proc. Natl. Acad. Sci.* **117**, 29330–29337 (2020).

105. Harper, N. S. & McAlpine, D. Optimal neural population coding of an auditory spatial cue. *Nature* **430**, 682–686 (2004).

106. Moiseff, A. & Konishi, M. Neuronal and behavioral sensitivity to binaural time differences in the owl. *J. Neurosci.* **1**, 40–48 (1981).

107. Carr, C. E. & Konishi, M. A circuit for detection of interaural time differences in the brain stem of the barn owl. *J. Neurosci.* **10**, 3227–3246 (1990).

108. Mackevicius, E. L., Best, M. D., Saal, H. P. & Bensmaia, S. J. Millisecond precision spike timing shapes tactile perception. *J. Neurosci.* **32**, 15309–15317 (2012).

109. Carney, L. H. A model for the responses of low-frequency auditory-nerve fibers in cat. *J. Acoust. Soc. Am.* **93**, 401–417 (1993).

110. Zhang, X., Heinz, M. G., Bruce, I. C. & Carney, L. H. A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *J. Acoust. Soc. Am.* **109**, 648–670 (2001).

111. Tan, Q. & Carney, L. H. A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide. *J. Acoust. Soc. Am.* **114**, 2007–2020 (2003).

112. Zilany, M. S. A. & Bruce, I. C. Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats. *J. Acoust. Soc. Am.* **122**, 402–417 (2007).

113. Zilany, M. S. A., Bruce, I. C., Nelson, P. C. & Carney, L. H. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *J. Acoust. Soc. Am.* **126**, 2390–2412 (2009).

114. Shera, C. A., Guinan, J. J. & Oxenham, A. J. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc. Natl. Acad. Sci.* **99**, 3318–3323 (2002).

115. Glasberg, B. & Moore, B. C. J. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* **47**, 103–138 (1990).

116. Baby, D., Van Den Broucke, A. & Verhulst, S. A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. *Nat. Mach. Intell.* **3**, 134–143 (2021).

117. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems* (eds. Pereira, F., Burges, C. J., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).

118. Yadav, S. & Foster, M. E. GISE-51: A scalable isolated sound events dataset. Preprint at https://doi.org/10.48550/arXiv.2103.12306 (2021).

119. Fonseca, E., Favory, X., Pons, J., Font, F. & Serra, X. FSD50k: An open dataset of human-labeled sound events. Preprint at https://doi.org/10.48550/arXiv.2010.00475 (2022).

120. Gemmeke, J. F. *et al.* Audio Set: An ontology and human-labeled dataset for audio events. in *Proc. IEEE ICASSP 2017* (New Orleans, LA, 2017).

121. Kell, A. J. E. & McDermott, J. H. Invariance to background noise as a signature of non-primary auditory cortex. *Nat. Commun.* **10**, 3958 (2019).

122. McWalter, R. & McDermott, J. H. Illusory sound texture reveals multi-second statistical completion in auditory scene analysis. *Nat. Commun.* **10**, 5096 (2019).

123. Shinn-Cunningham, B. G., Desloge, J. G. & Kopco, N. 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)* 183–186 (2001). doi:10.1109/ASPAA.2001.969573.

124. Kawahara, H. *et al.* Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *2008 IEEE Int. Conf. Acoust. Speech Signal Process.* 3933–3936 (2008) doi:10.1109/icassp.2008.4518514.

125. Algazi, V. R., Duda, R. O., Thompson, D. M. & Avendano, C. The CIPIC HRTF database. in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)* 99–102 (2001). doi:10.1109/ASPAA.2001.969552.

126. Litovsky, R. Y. & Godar, S. P. Difference in precedence effect between children and adults signifies development of sound localization abilities in complex listening tasks. *J. Acoust. Soc. Am.* **128**, 1979–1991 (2010).

127. Litovsky, R. Y., Colburn, H. S., Yost, W. A. & Guzman, S. J. The precedence effect. *J. Acoust. Soc. Am.* **106**, 1633–1654 (1999).

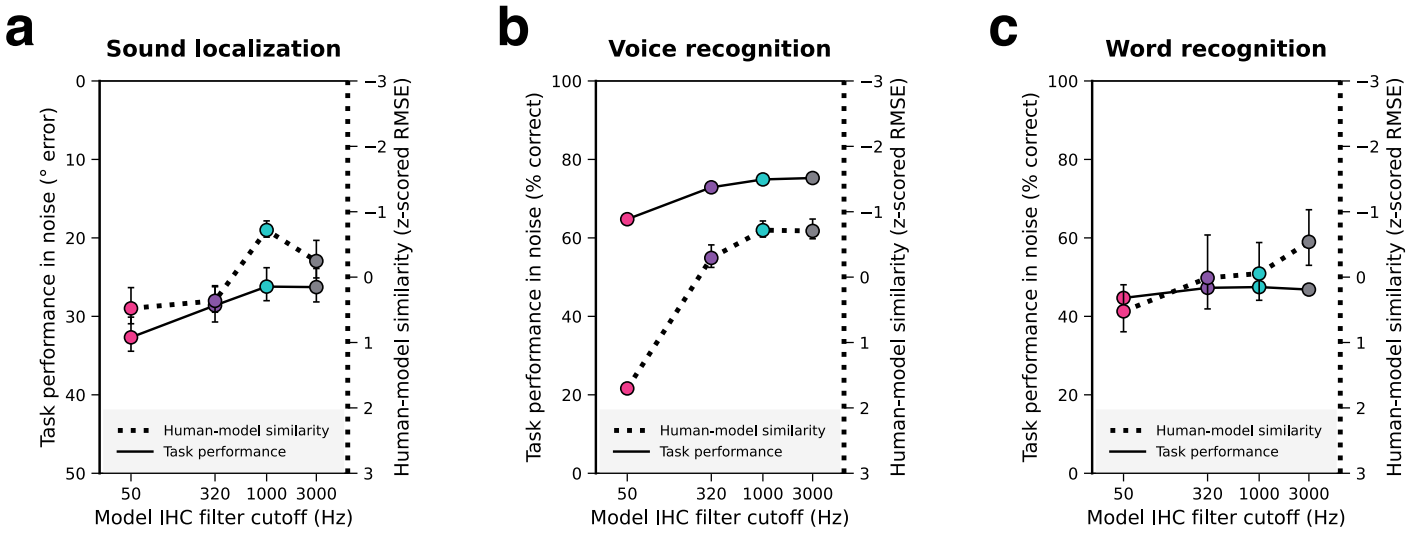
128. Giannoulis, D. *et al.* A database and challenge for acoustic scene classification and event detection. in *21st European Signal Processing Conference (EUSIPCO 2013)* 1–5 (2013).

129. Ardila, R. *et al.* Common voice: A massively-multilingual speech corpus. *ArXiv Prepr. ArXiv191206670* (2019).

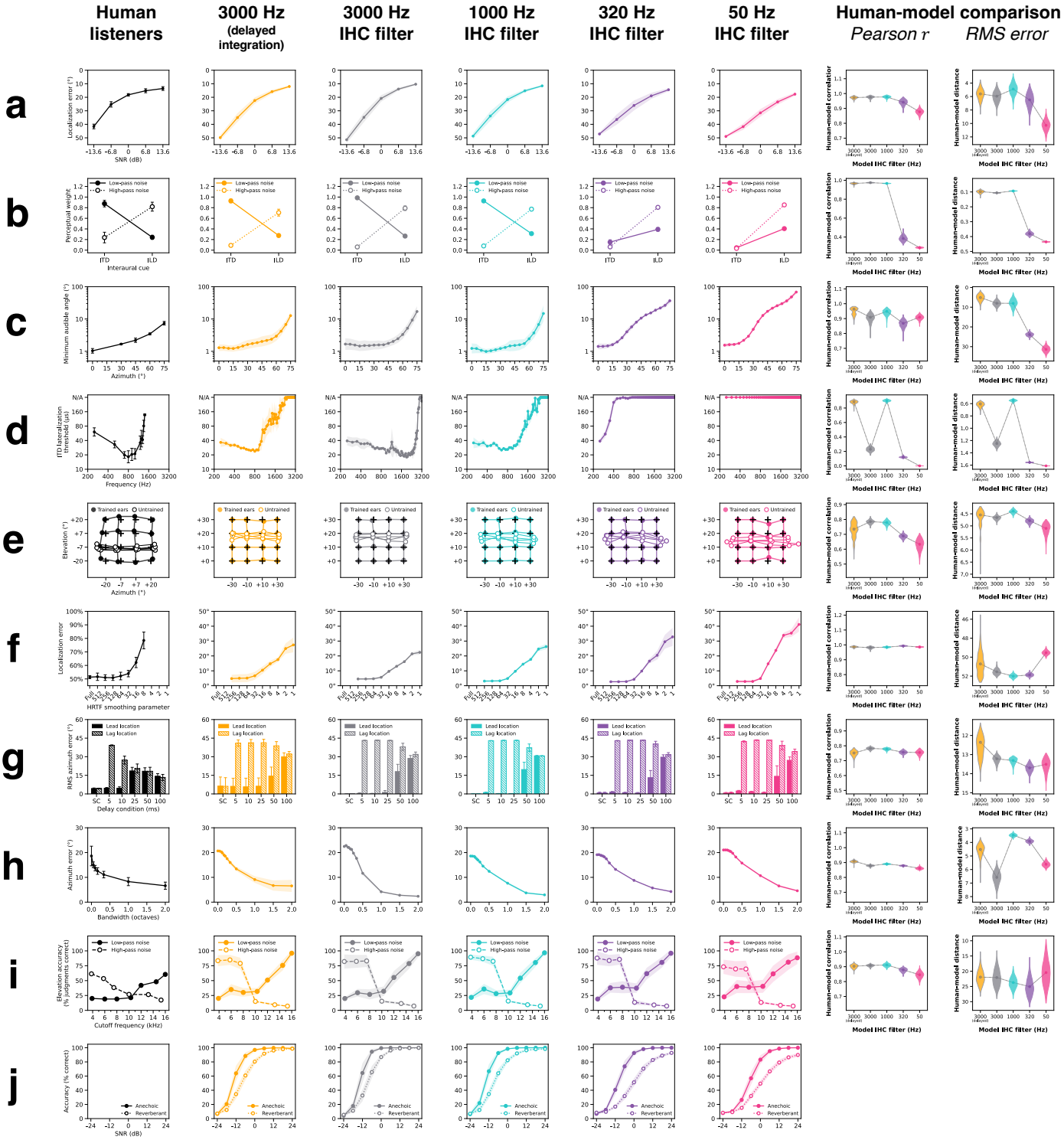
130. Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I. & Bittner, R. MUSDB18-a corpus for music separation. (2017).

131. Woods, K. J. P., Siegel, M. H., Traer, J. & McDermott, J. H. Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* **79**, 2064–2072 (2017).

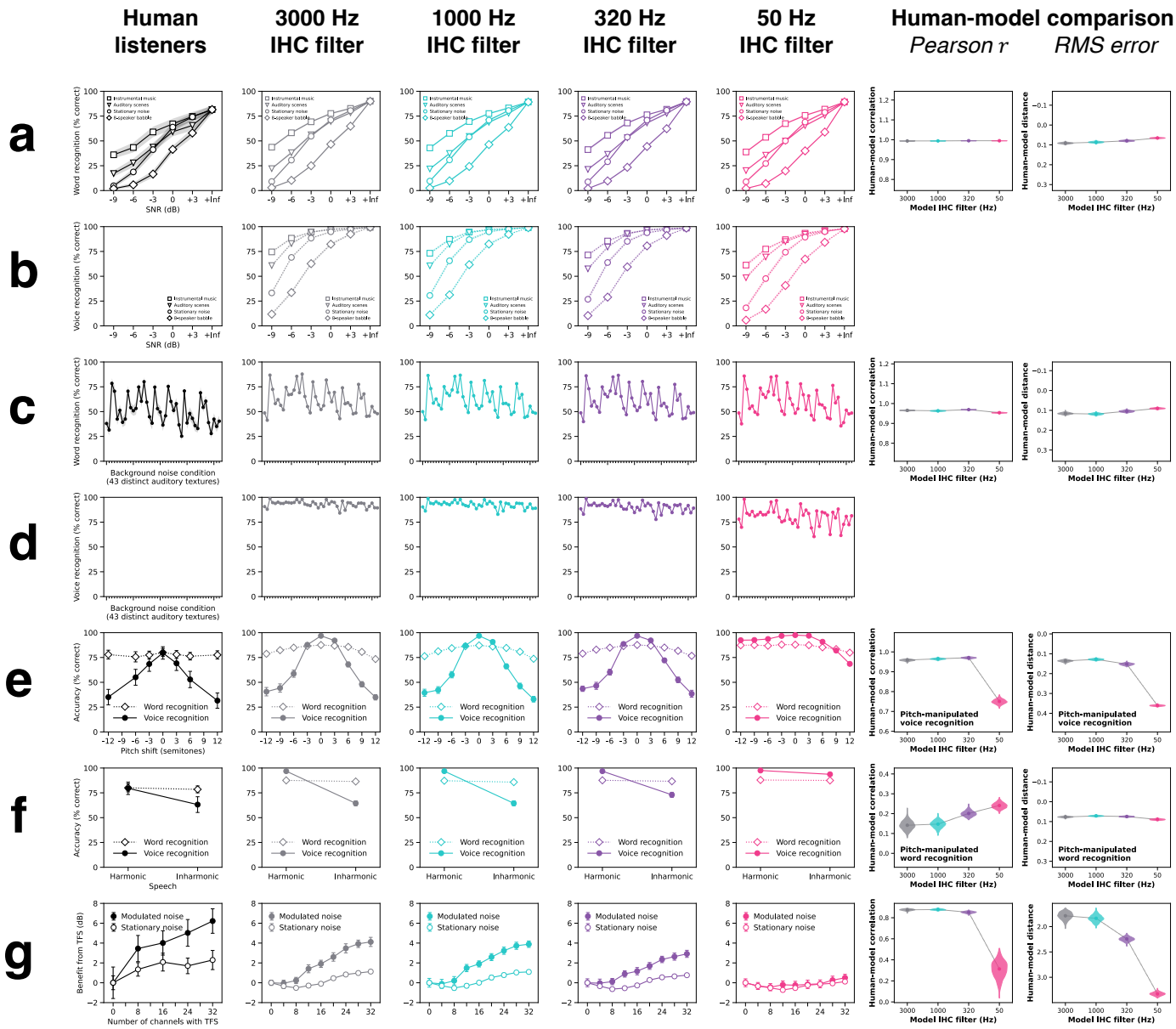
SUPPLEMENTARY INFORMATION



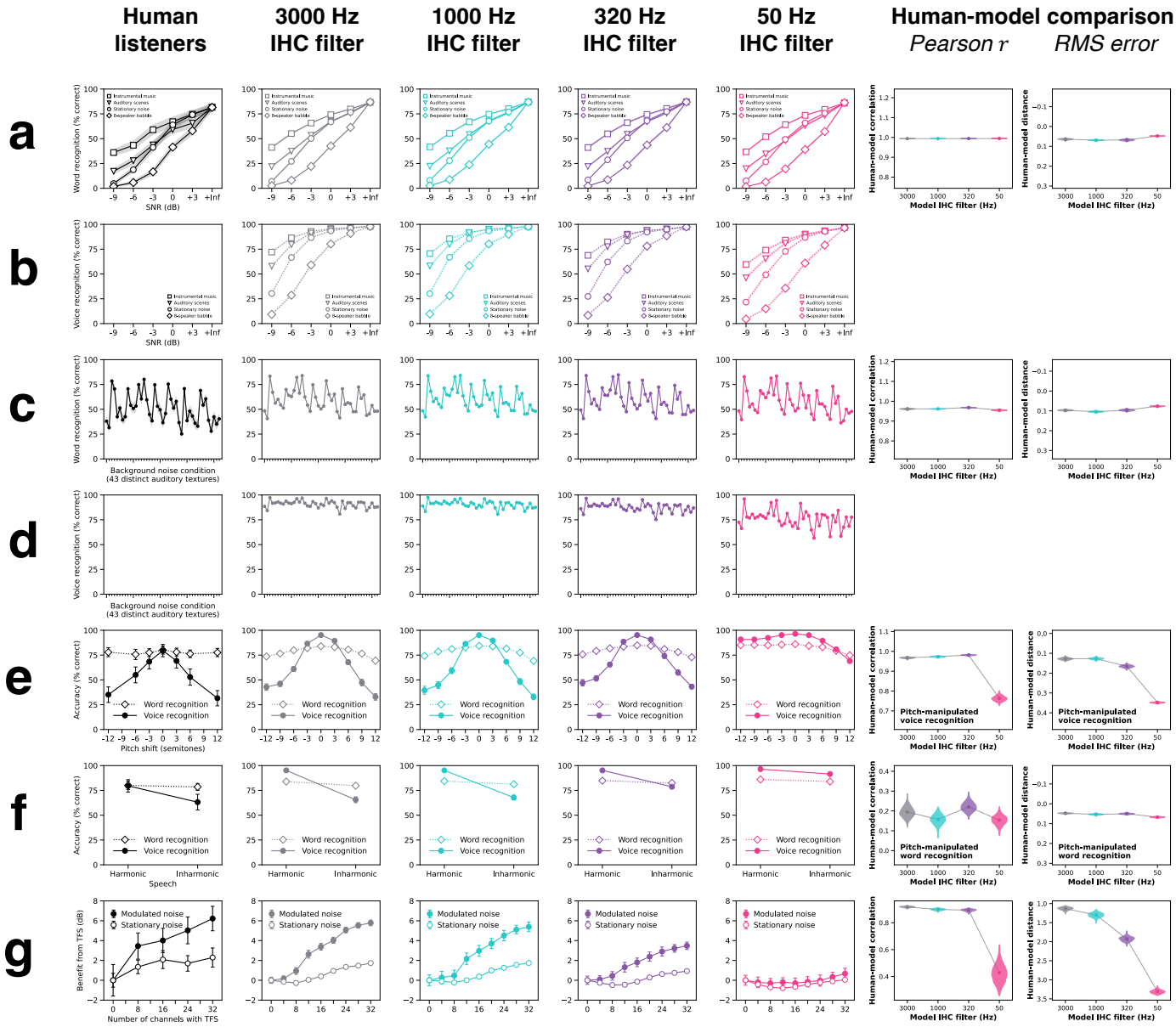
Supplementary Fig. 1 | Models with access to phase-locked spike timing have better and more human-like hearing (alternate human-model similarity metric). As in Fig. 2, each panel corresponds to a different task and summarizes the effect of auditory nerve phase locking limit on i) naturalistic model task performance and ii) overall human-model behavioral similarity. The difference from Fig. 2 is that overall human-model behavioral similarity is quantified as the root-mean-squared error between analogous human and model data points, z-scored and then averaged across all experiments for each model task (right y-axes; dotted lines). The right y-axes are inverted such that higher positions correspond to more human-like behavior. Naturalistic task performance is quantified as a single number averaged across noise conditions (left y-axes; solid lines). Error bars indicate 95% confidence intervals bootstrapped across 10 network architectures for each model. **a.** Sound localization. The left y-axis plots mean absolute error for the sound localization model and is inverted so that better model performance corresponds to higher positions on the y-axis. **b.** Voice recognition. Here and in **c,** the left y-axes plot percent correct for the model when tested on speech in noise. **c.** Word recognition.



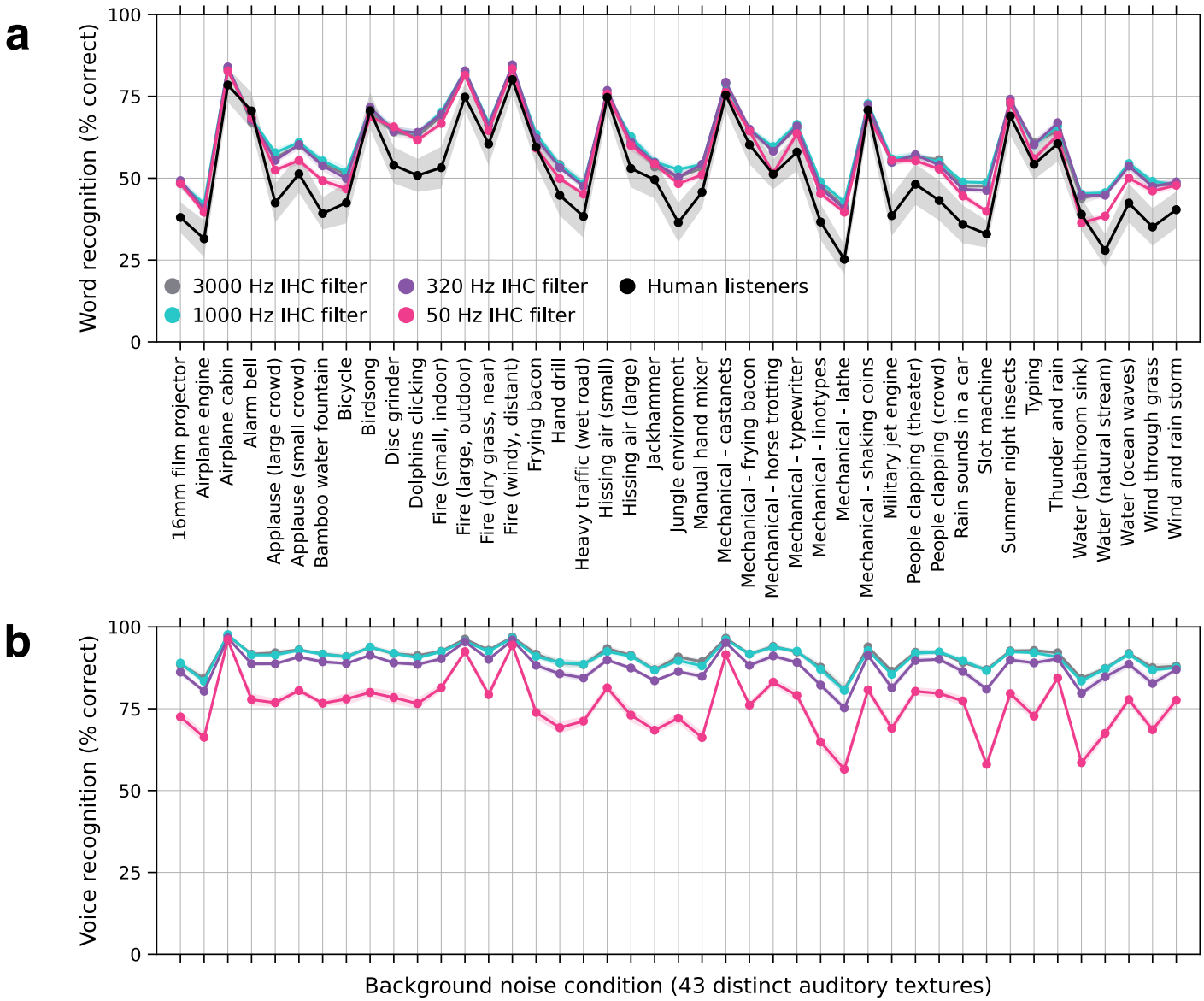
Supplementary Fig. 2 | Effect of phase locking on all sound localization experiments. This grid summarizes the behavioral data used to measure human-model similarity scores for the localization models. The first six columns correspond to human listeners and models optimized with different phase locking limits. The second (orange) column corresponds to the 3000 Hz phase locking model with network architectures modified to delay binaural integration. The rightmost two columns quantify human-similarity by measuring Pearson correlations and root-mean-squared error between analogous human and model data points. Violin plots depict bootstrapped distributions of human-model similarity scores across 10 network architectures per phase locking condition. Rows correspond to 10 different sound localization experiments. **a.** Sound localization in noise. **b.** Minimum audible angle vs. frequency. **c.** ITD / ILD cue weighting. **d.** ITD lateralization vs. frequency. **e.** Effect of changing ears. **f.** Effect of smoothing spectral cues. **g.** Precedence effect. **h.** Bandwidth dependency of localization. **i.** Median plane spectral cues. **j.** Speech localization in noise and reverberation (model experiment only). All model error bars indicate ±2 standard errors of the mean across 10 network architectures.



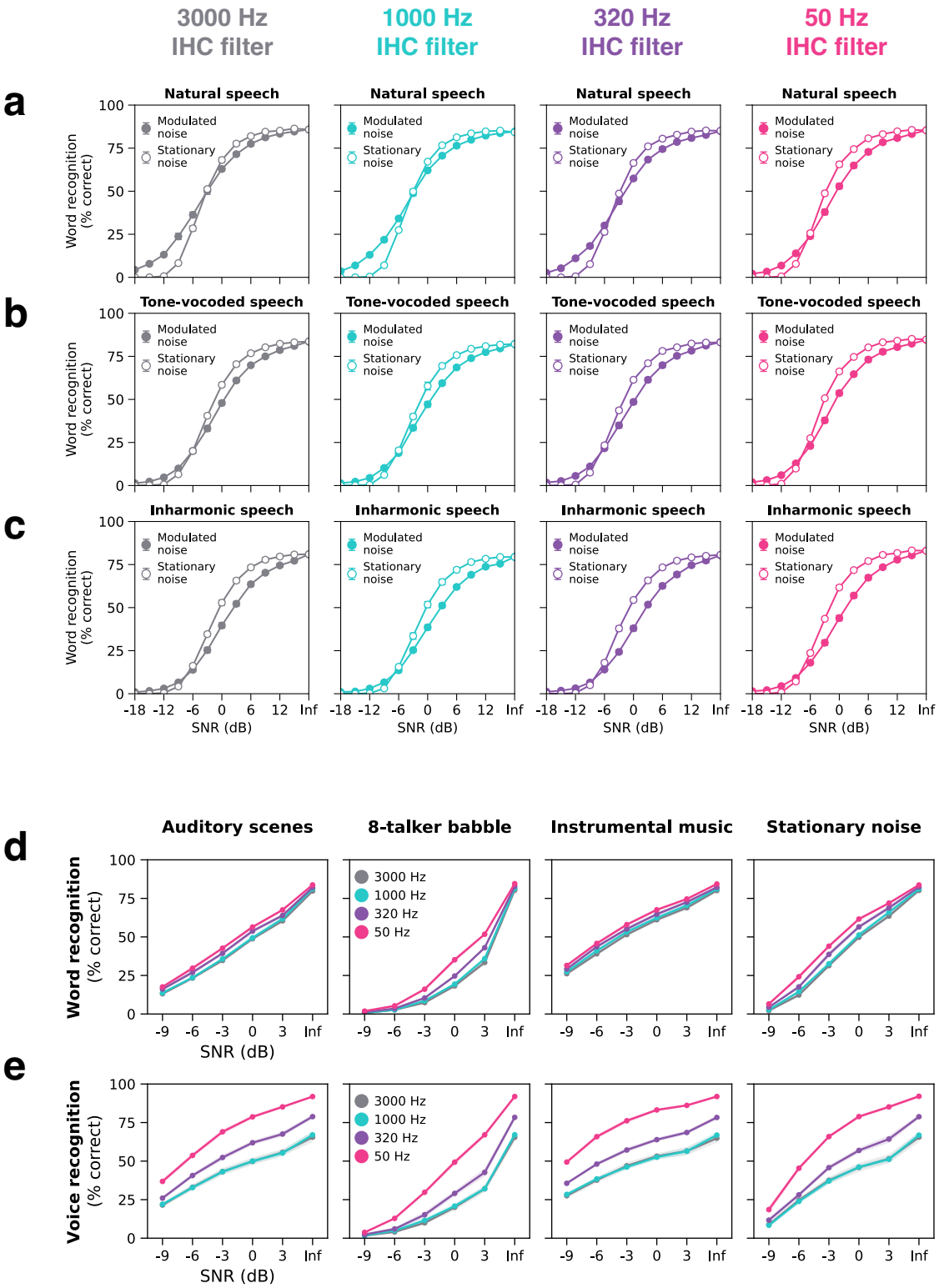
Supplementary Fig. 3 | Models optimized separately for word and voice recognition -- effect of phase locking on all speech experiments. The same network architectures optimized jointly for word and voice recognition in the main text were also optimized separately for the word and voice recognition tasks for each phase locking condition. This yielded similar results to the jointly optimized models. The first five columns correspond to human listeners and models optimized with different phase locking limits. The rightmost two columns quantify human-similarity by measuring Pearson correlations and root-mean-squared error between analogous human and model data points. Violin plots depict bootstrapped distributions of human-model similarity scores across 10 network architectures per phase locking condition. Rows correspond to 7 speech experiments. **a.** Word recognition in real-world noise conditions. **b.** Voice recognition in real-world noise conditions (model experiment only). **c.** Word recognition in 43 distinct auditory textures at -3 dB SNR. **d.** Voice recognition in 43 distinct auditory textures at -3 dB SNR (model experiment only). **e.** Word and voice recognition with pitch-shifted speech. **f.** Word and voice recognition with harmonic and inharmonic speech. **g.** Effect of tone vocoding on word recognition in stationary and modulated noise. All model error bars indicate ±2 standard errors of the mean across 10 network architectures.



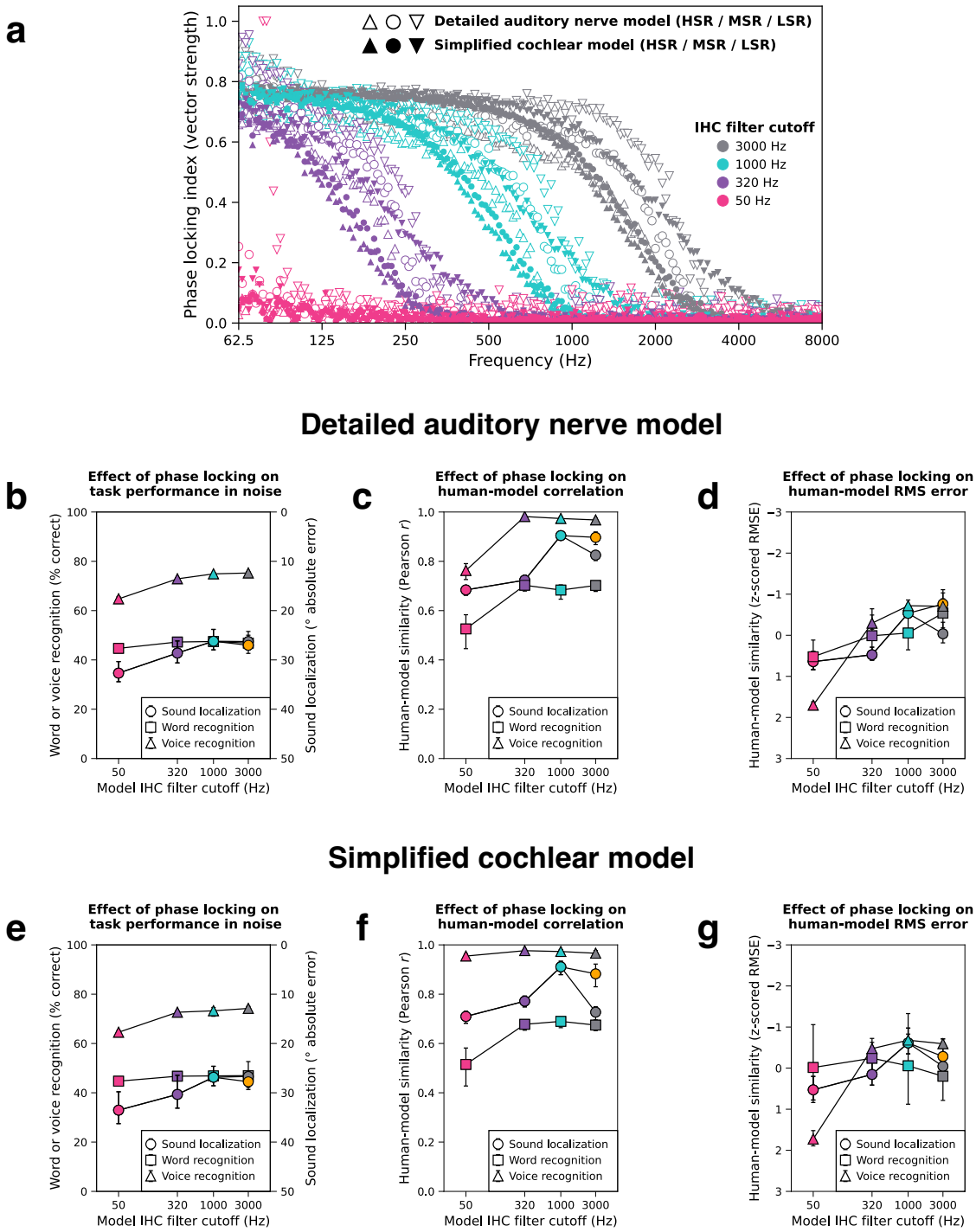
Supplementary Fig. 4 | Models optimized jointly for word and voice recognition -- effect of phase locking on all speech experiments. This grid summarizes the behavioral data used to measure human-model similarity scores for the word and voice recognition models. The first five columns correspond to human listeners and models optimized with different phase locking limits. The rightmost two columns quantify human-similarity by measuring Pearson correlations and root-mean-squared error between analogous human and model data points. Violin plots depict bootstrapped distributions of human-model similarity scores across 10 network architectures per phase locking condition. Rows correspond to 7 speech experiments. **a.** Word recognition in real-world noise conditions. **b.** Voice recognition in real-world noise conditions (model experiment only). **c.** Word recognition in 43 distinct auditory textures at -3 dB SNR. **d.** Voice recognition in 43 distinct auditory textures at -3 dB SNR (model experiment only). **e.** Word and voice recognition with pitch-shifted speech. **f.** Word and voice recognition with harmonic and inharmonic speech. **g.** Effect of tone vocoding on word recognition in stationary and modulated noise. All model error bars indicate ±2 standard errors of the mean across 10 network architectures.



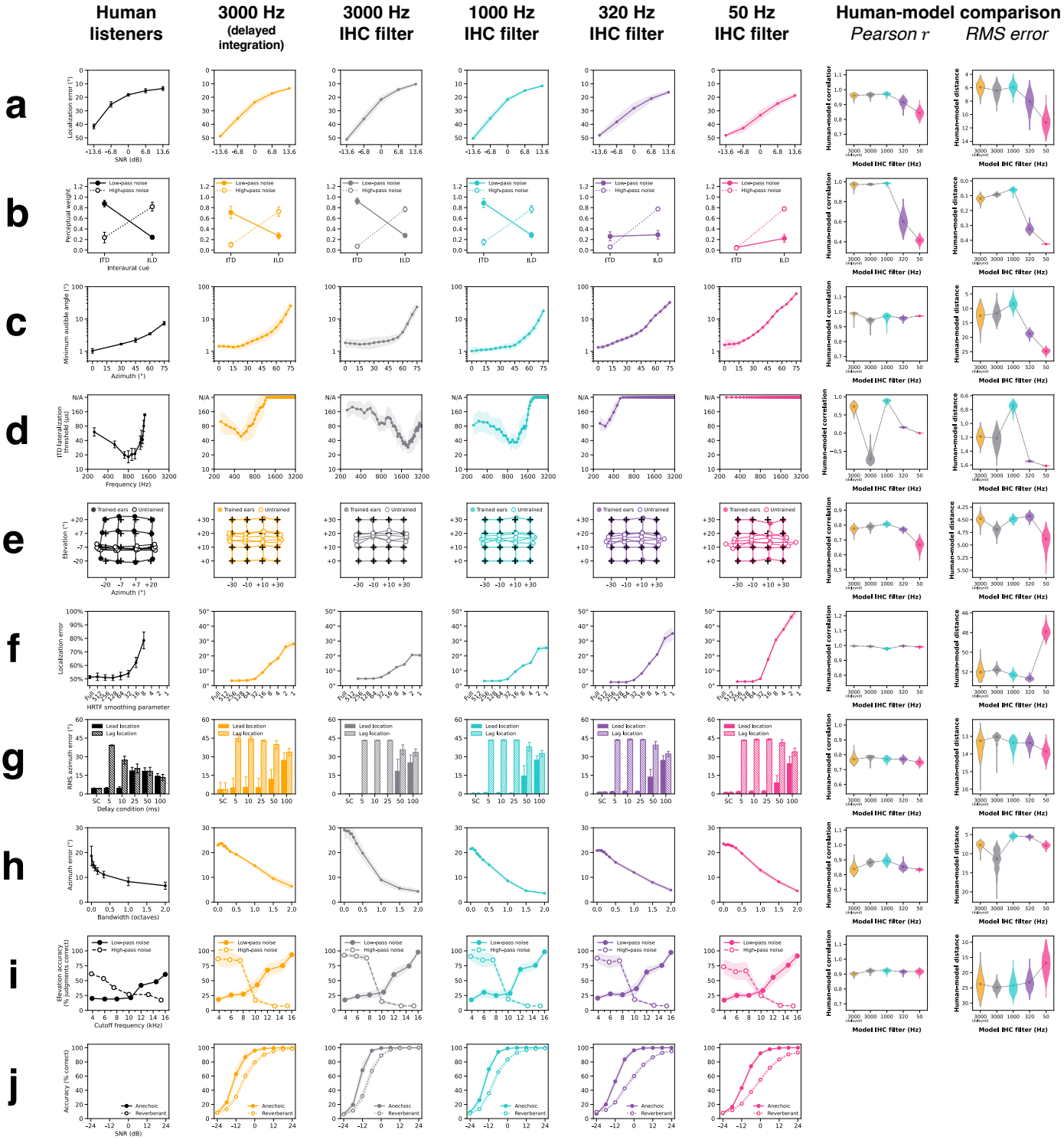
Supplementary Fig. 5 | Word and voice recognition in real-world auditory textures. a. Human and model word recognition for speech embedded in 43 distinct auditory textures at -3 dB SNR. **b.** Model voice recognition for the same stimuli. Error bars indicate ±2 standard errors of the mean across human participants or network architectures.



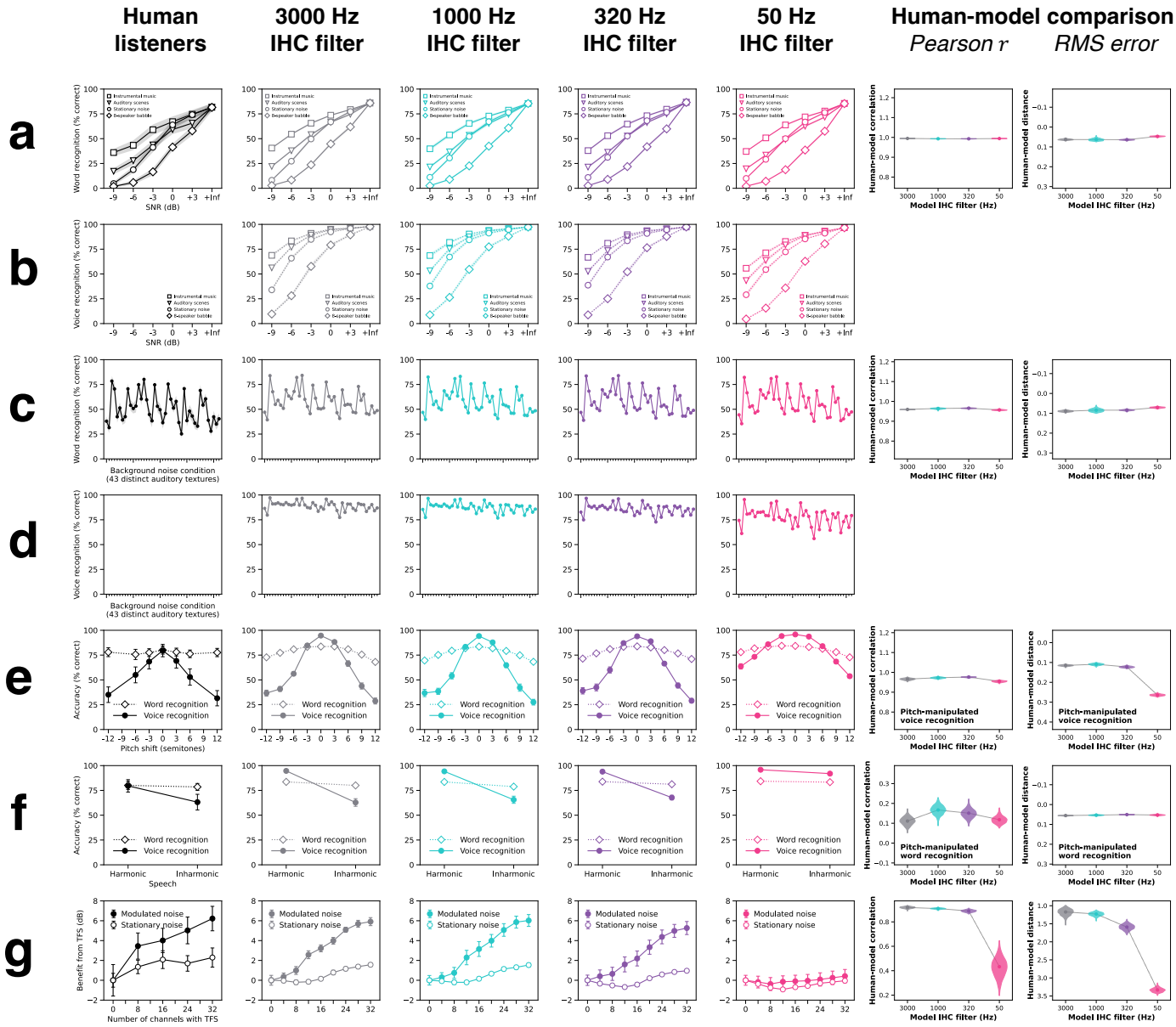
Supplementary Fig. 6 | Model word and voice recognition with inharmonic speech in noise. To directly compare effects of the inharmonicity and tone vocoding stimulus manipulations on model word recognition in noise, we measured word recognition accuracy in stationary and modulated speech-shaped noise at SNRs between -18 and +15 dB in 3 dB increments using (**a.**) natural, (**b.**) tone-vocoded, and (**c.**) inharmonic versions of the same speech. The tone-vocoded speech was fully vocoded (0 channels with intact TFS). **d.** Model word recognition with inharmonic speech as a function of SNR in four different types of real-world noise. **e.** Model voice recognition with inharmonic speech as a function of SNR in four different types of real-world noise. All error bars indicate ±2 standard errors of the mean across 10 network architectures.



Supplementary Fig. 7 | Comparison of model results with detailed vs. simplified cochlear stages. a. The strength of phase locking as a function of frequency for simulated auditory nerve fibers under four inner hair cell (IHC) low-pass filter cutoffs (different colors). Nerve fibers simulated with the detailed auditory nerve model[46] (open symbols) and the simplified cochlear model (solid symbols) exhibit similar roll-offs in phase locking. The three different symbol shapes indicate high-, medium-, and low-spontaneous-rate (HSR, MSR, and LSR) auditory nerve fibers. Panels **b**, **c**, and **d** present results for models operating on input from the detailed auditory nerve model. **b.** Aggregate measures of task performance in noise as a function of phase locking. Word and voice recognition performance are plotted on the left y-axis (solid lines). Localization model performance is plotted on the right y-axis (dotted lines). **c.** Aggregate measure of human-model similarity (quantified as the Pearson correlation coefficient averaged across all experiments for each task) as a function of phase locking. **d.** Aggregate measure of human-model similarity (quantified as the z-scored root-mean-squared error averaged across all experiments for each task) as a function of phase locking. Panels **e**, **f**, and **g** are formatted identically to **b**, **c**, and **d** but present results for models operating on input from the simplified cochlear model. The orange symbol in panels **b** - **g** represents the 3000 Hz sound localization model with delayed interaural integration (see Fig. 4). Error bars indicate 95% confidence intervals bootstrapped across 10 network architectures for each model.



Supplementary Fig. 8 | Simplified cochlear model -- effect of phase locking on all localization experiments. This grid summarizes the behavioral data used to measure human-model similarity scores for localization models with the simplified cochlear stage. The first six columns correspond to human listeners and models optimized with different phase locking limits. The second (orange) column corresponds to the 3000 Hz phase locking model with network architectures modified to delay binaural integration. The rightmost two columns quantify human-similarity by measuring Pearson correlations and root-mean-squared error between analogous human and model data points. Violin plots depict bootstrapped distributions of human-model similarity scores across 10 network architectures per phase locking condition. Rows correspond to 10 different sound localization experiments. **a.** Sound localization in noise. **b.** Minimum audible angle vs. frequency. **c.** ITD / ILD cue weighting. **d.** ITD lateralization vs. frequency. **e.** Effect of changing ears. **f.** Effect of smoothing spectral cues. **g.** Precedence effect. **h.** Bandwidth dependency of localization. **i.** Median plane spectral cues. **j.** Speech localization in noise and reverberation (model experiment only). All model error bars indicate ±2 standard errors of the mean across 10 network architectures.



Supplementary Fig. 9 | Simplified cochlear model -- effect of phase locking on all speech experiments. This grid summarizes the behavioral data used to measure human-model similarity scores for the word and voice recognition models with the simplified cochlear stage. The first five columns correspond to human listeners and models optimized with different phase locking limits. The rightmost two columns quantify human-similarity by measuring Pearson correlations and root-mean-squared error between analogous human and model data points. Violin plots depict bootstrapped distributions of human-model similarity scores across 10 network architectures per phase locking condition. Rows correspond to 7 speech experiments. **a.** Word recognition in real-world noise conditions. **b.** Voice recognition in real-world noise conditions (model experiment only). **c.** Word recognition in 43 distinct auditory textures at -3 dB SNR. **d.** Voice recognition in 43 distinct auditory textures at -3 dB SNR (model experiment only). **e.** Word and voice recognition with pitch-shifted speech. **f.** Word and voice recognition with harmonic and inharmonic speech. **g.** Effect of tone vocoding on word recognition in stationary and modulated noise. All model error bars indicate ±2 standard errors of the mean across 10 network architectures.

Architecture	arch_01	arch_02	arch_03	arch_04	arch_05	arch_06	arch_07	arch_08	arch_09	arch_10
Operation	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]	input [50,10000,6]
1	conv0 [1, 8, 32]	conv0 [2, 8, 32]	conv0 [1, 4, 32]	conv0 [3, 8, 32]	conv0 [2, 32, 32]	conv0 [1, 64, 32]	conv0 [1, 16, 32]	conv0 [1, 64, 32]	conv0 [3, 32, 32]	conv0 [2, 4, 32]
2	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 2]	mpool0 [1, 8]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [1, 1]	mpool0 [2, 2]
3	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0
4	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0	bnorm0
5	conv1 [1, 64, 32]	conv1 [3, 16, 32]	conv1 [3, 32, 32]	conv1 [3, 8, 32]	conv1 [1, 4, 64]	conv1 [2, 4, 64]	conv1 [1, 8, 32]	conv1 [2, 16, 32]	conv1 [2, 16, 32]	conv1 [2, 4, 32]
6	mpool1 [1, 1]	mpool1 [1, 1]	mpool1 [1, 8]	mpool1 [1, 2]	mpool1 [1, 4]	mpool1 [1, 1]	mpool1 [1, 2]	mpool1 [1, 8]	mpool1 [1, 4]	mpool1 [1, 4]
7	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1
8	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1	bnorm1
9	conv2 [1, 64, 32]	conv2 [2, 4, 32]	conv2 [3, 32, 64]	conv2 [1, 32, 64]	conv2 [3, 2, 64]	conv2 [1, 32, 64]	conv2 [2, 4, 64]	conv2 [2, 4, 64]	conv2 [2, 32, 64]	conv2 [3, 16, 64]
10	mpool2 [1, 8]	mpool2 [1, 8]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [2, 4]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [1, 1]	mpool2 [1, 2]
11	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2
12	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2	bnorm2
13	conv3 [2, 4, 64]	conv3 [3, 16, 64]	conv3 [1, 8, 64]	conv3 [3, 8, 64]	conv3 [2, 8, 64]	conv3 [3, 4, 128]	conv3 [2, 32, 64]	conv3 [2, 16, 64]	conv3 [3, 4, 64]	conv3 [1, 2, 128]
14	mpool3 [2, 4]	mpool3 [1, 1]	mpool3 [1, 4]	mpool3 [2, 4]	mpool3 [1, 1]	mpool3 [1, 1]	mpool3 [1, 4]	mpool3 [1, 1]	mpool3 [1, 4]	mpool3 [1, 2]
15	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3
16	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3	bnorm3
17	conv4 [3, 8, 128]	conv4 [1, 8, 64]	conv4 [3, 8, 64]	conv4 [2, 2, 128]	conv4 [1, 16, 64]	conv4 [2, 16, 128]	conv4 [3, 2, 64]	conv4 [1, 16, 64]	conv4 [3, 8, 128]	flatten
18	mpool4 [1, 1]	mpool4 [1, 4]	mpool4 [1, 1]	mpool4 [1, 4]	mpool4 [1, 4]	mpool4 [1, 2]	mpool4 [1, 1]	mpool4 [1, 2]	mpool4 [1, 4]	fc0 [512]
19	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu_fc0
20	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	bnorm4	norm_fc0
21	conv5 [3, 32, 128]	conv5 [3, 8, 128]	conv5 [1, 2, 64]	conv5 [1, 4, 256]	conv5 [3, 4, 128]	conv5 [1, 2, 256]	conv5 [1, 2, 64]	conv5 [2, 32, 128]	conv5 [3, 2, 256]	dropout
22	mpool5 [1, 4]	mpool5 [1, 4]	mpool5 [1, 1]	mpool5 [1, 1]	mpool5 [1, 2]	mpool5 [1, 1]	mpool5 [2, 4]	mpool5 [1, 4]	mpool5 [1, 2]	fc [504]
23	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	
24	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	bnorm5	
25	conv6 [3, 4, 256]	conv6 [2, 2, 128]	conv6 [2, 2, 64]	conv6 [3, 2, 256]	conv6 [3, 4, 256]	conv6 [3, 4, 256]	conv6 [1, 8, 128]	conv6 [2, 16, 128]	conv6 [2, 8, 512]	
26	mpool6 [1, 1]	mpool6 [1, 2]	mpool6 [2, 4]	mpool6 [1, 1]	mpool6 [1, 1]	mpool6 [1, 2]	mpool6 [1, 1]	mpool6 [1, 1]	mpool6 [1, 1]	
27	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	
28	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	bnorm6	
29	conv7 [3, 8, 256]	conv7 [3, 2, 256]	conv7 [2, 4, 128]	conv7 [2, 2, 256]	conv7 [3, 4, 256]	flatten	flatten	conv7 [1, 2, 128]	conv7 [3, 4, 512]	
30	mpool7 [1, 2]	mpool7 [1, 2]	mpool7 [1, 1]	mpool7 [1, 2]	mpool7 [1, 1]	fc0 [512]	fc0 [512]	mpool7 [1, 1]	mpool7 [1, 2]	
31	relu7	relu7	relu7	relu7	relu7	relu_fc0	relu_fc0	relu7	relu7	
32	bnorm7	bnorm7	bnorm7	bnorm7	bnorm7	norm_fc0	norm_fc0	bnorm7	bnorm7	
33	flatten	conv8 [1, 8, 512]	conv8 [1, 8, 128]	flatten	conv8 [2, 4, 256]	dropout	dropout	conv8 [3, 16, 128]	conv8 [1, 3, 512]	
34	fc0 [512]	mpool8 [1, 2]	mpool8 [1, 1]	fc0 [512]	mpool8 [1, 2]	fc [504]	fc [504]	mpool8 [1, 4]	mpool8 [1, 1]	
35	relu_fc0	relu8	relu8	relu_fc0	relu8			relu8	relu8	
36	norm_fc0	bnorm8	bnorm8	norm_fc0	bnorm8			bnorm8	bnorm8	
37	dropout	flatten	conv9 [3, 2, 128]	dropout	flatten			flatten	flatten	
38	fc [504]	fc0 [512]	mpool9 [1, 4]	fc [504]	fc0 [512]			fc0 [512]	fc0 [512]	
39		relu_fc0	relu9		relu_fc0			relu_fc0	relu_fc0	
40		norm_fc0	bnorm9		norm_fc0			norm_fc0	norm_fc0	
41		dropout	flatten		dropout			dropout	dropout	
42		fc [504]	fc0 [512]		fc [504]			fc [504]	fc [504]	
43			relu_fc0							
44			norm_fc0							
45			dropout							
46			fc [504]							
47										

Supplementary Table 1 | Neural network architectures for sound localization models. Grey bands indicate blocks of convolution, pooling, nonlinear rectification, and normalization operations. The convolution operations highlighted in orange were replaced with grouped convolutions (2 groups for the left and right ear) when network architectures were modified to delay binaural integration. Legend:

- $conv\ [h, w, k]$: convolutional layer with h = kernel height (frequency dimension), w = kernel width (time dimension), and k = number of kernels
- $relu$: rectified linear unit activation function
- $mpool\ [s_f, s_t]$: max pooling operation with stride s_f in the frequency dimension and stride s_t in the time dimension
- $bnorm$: batch normalization operation
- $flatten$: multidimensional representation reshaped to a vector
- $fc\ [N]$: fully-connected layer with N units
- $dropout$: dropout regularization with 50% dropout rate

Architecture	arch0_0000	arch0_0001	arch0_0002	arch0_0004	arch0_0006	arch0_0007	arch0_0008	arch0_0009	arch0_0016	arch0_0017
Operation	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]	input [50, 20000, 3]
1	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm	input_lnorm
2	conv0 [2, 42, 32]	conv0 [1, 84, 32]	conv0 [4, 21, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]	conv0 [2, 42, 32]
3	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0	relu0
4	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]	hpool0 [2, 4]
5	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0	lnorm0
6	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [4, 9, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]	conv1 [2, 18, 64]
7	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1	relu1
8	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]	hpool1 [2, 4]
9	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1	lnorm1
10	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [12, 3, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]	conv2 [6, 6, 128]
11	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2	relu2
12	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]	hpool2 [1, 4]
13	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2	lnorm2
14	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [3, 12, 256]	conv3 [12, 3, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]	conv3 [6, 6, 256]
15	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3	relu3
16	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]	hpool3 [1, 4]
17	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3	lnorm3
18	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]	conv4 [4, 16, 512]	conv4 [8, 8, 512]	conv4 [8, 8, 512]
19	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4	relu4
20	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]	hpool4 [1, 1]
21	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4	lnorm4
22	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]	conv5 [6, 6, 512]
23	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5	relu5
24	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]	hpool5 [1, 1]
25	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5	lnorm5
26	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]	conv6 [8, 8, 512]
27	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6	relu6
28	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]	hpool6 [2, 4]
29	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6	lnorm6
30	flatten	flatten	flatten	flatten	flatten	flatten	flatten	flatten	conv7 [2, 8, 512]	conv7 [8, 2, 512]
31	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	fc0 [512]	relu7	relu7
32	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	relu_fc0	hpool7 [1, 1]	hpool7 [1, 1]
33	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	norm_fc0	lnorm7	lnorm7
34	dropout	dropout	dropout	dropout	dropout	dropout	dropout	dropout	flatten	flatten
35	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc [433, 794]	fc0 [512]	fc0 [512]
36									relu_fc0	relu_fc0
37									norm_fc0	norm_fc0
38									dropout	dropout
39									fc [433, 794]	fc [433, 794]
40										

Supplementary Table 2 | Neural network architectures for word and voice recognition models. Grey bands indicate blocks of convolution, pooling, nonlinear rectification, and normalization operations. For networks jointly optimized for word and voice recognition, there were two fully-connected read-out layers in parallel, one for each task (433 units for voice recognition and 794 units for word recognition). Legend:

- $conv\ [h, w, k]$: convolutional layer with $h =$ kernel height (frequency dimension), $w =$ kernel width (time dimension), and $k =$ number of kernels
- $relu$: rectified linear unit activation function
- $hpool\ [s_f, s_t]$: Hanning window weighted averaged pooling operation with stride s_f in the frequency dimension and stride s_t in the time dimension
- $lnorm$: layer normalization operation
- $flatten$: multidimensional representation reshaped to a vector
- $fc\ [N]$: fully-connected layer with N units
- $fc\ [N_{voice}, N_{word}]$: two parallel fully-connected layers operating on the same input, one with N_{voice} units and one with N_{word} units
- $dropout$: dropout regularization with 50% dropout rate