Optimized feature gains explain and predict successes and failures of human selective listening

Ian M. Griffith[1,2,3], R. Preston Hess[1,2], 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]Program in Speech and Hearing Biosciences and Technology, Harvard, Cambridge, MA, USA
[4]Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA

Attention facilitates communication by enabling selective listening to sound sources of interest. However, little is known about why attentional selection succeeds in some conditions but fails in others. While neurophysiology implicates multiplicative feature gains in selective attention, it is unclear whether such gains can explain real-world attention-driven behavior. To investigate these issues, we optimized an artificial neural network with stimulus-computable, feature-based gains to recognize a cued talker's speech from binaural audio in "cocktail party" scenarios. Though not trained to mimic humans, the model matched human performance across diverse real-world conditions, exhibiting selection based both on voice qualities and spatial location. It also predicted novel attentional effects that we confirmed in human experiments, and exhibited signatures of "late selection" like those seen in human auditory cortex. The results suggest that human-like attentional strategies naturally arise from optimization of feature gains for selective listening, offering a normative account of the mechanisms—and limitations—of auditory attention.

Introduction

Organisms often base behavior on one of many objects in their environment. This ability typically involves endogenous ("top-down") attention – a change in internal state that allows an organism to willfully "select" an object of interest for further processing, memory, or a behavioral response. Selective attention has been a central focus of cognitive science and neuroscience since the 1950s[1], and much is known about both attentional abilities and their neural correlates[2-8].

Neurophysiological observations suggest a mechanistic account of attentional selection as multiplicative gains that enhance the perceptual features of an attended object. Specifically, attention to particular features tends to cause the responses of neurons tuned to those features to be scaled upwards, enhancing the representation of objects containing those features[9-12]. Other types of effects on neural tuning[12,13] can be explained by combining multiplicative gains with normalization[14]. Consistent with such findings, human neuroscience studies show attentional enhancement of attended sound sources[15-20]. However, it remains unclear whether feature-based multiplicative gains are sufficient to account for real-world attention-based abilities, in part because we have lacked working models of attention that can be evaluated in everyday settings. In particular, although computational models of sensory systems have made notable advances in being able to account for some types of human judgments of images and sounds[21,22], they have thus far largely not incorporated attentional mechanisms (see Discussion).

Speech comprehension is one setting in which selective attention is essential in daily life. We routinely must understand what someone is saying despite the presence of other concurrent talkers (the "cocktail party problem")[1,23,24]. Human attentional selection of one voice among others has been demonstrated in many settings[25-28] and is known to depend on the features of individual sound sources – both spatial locations and voice qualities such as pitch. However, human attentional selection can be prone to failure in some settings, as when a distractor voice shares features with the target voice[26,29]. The reasons for these failures remain poorly understood. For instance, it has been unclear whether human selection errors represent suboptimal strategies, or whether such errors are largely inevitable given the feature overlap between human voices.

We used the cocktail party problem[23,24] as a setting in which to explore computational accounts of feature-based attention. We tested whether a task-optimized model equipped with multiplicative gains applied to sensory representations would replicate human selective listening behavior. We assumed gains to be functions of a memory of the attentional target's properties, and used a task in which the target properties could be estimated from prior exposure to a target talker's voice. We found that models with such multiplicative gains that were optimized to recognize the words of a cued talker reproduced a wide range of characteristics of human auditory attention, including its sensitivity to both spatial and nonspatial aspects of sound, and occasional failures of selection. They also predicted two previously undocumented traits of human attentional selection that we subsequently confirmed experimentally.

The results indicate that multiplicative gains, when combined with standard filtering, pooling, and normalization operations, are sufficient to account for both the successes and failures of human feature-based attention, and help explain why attentional effects are most evident in later stages of sensory systems. The framework we propose is general, and should be applicable to any behavior involving feature-based attention.

Results

Feature-based attention task

To study feature-based attention, we used a task in which a listener first heard an excerpt of a target voice (the "cue"), then heard a mixture of the target voice with a distractor voice, and then reported the middle word spoken by the target voice (Fig. 1a). The excerpt of the target voice used for the mixture was different from that used for the cue, such that the cue indicated the target talker's voice and spatial location, but not the words spoken by the target talker in the mixture. To solve the task, the listener had to use the properties of the cue (it's voice properties and/or spatial position) to select the target voice from the mixture. Depending on the experiment, sounds could be presented over headphones, with the same audio signal to both ears (eliminating any spatial cues), or from different spatial locations (with each ear receiving a different audio signal).

Model optimization

We used supervised deep learning to build a model that could perform this task. Models were trained to report the middle word in a 2-second speech excerpt spoken by a cued talker within a mixture of talkers. Both cue and mixture were presented as binaural audio. During training, audio was spatially rendered at locations within simulated reverberant rooms. To emulate the variety of scenarios humans encounter in the real-world, each mixture in the training data varied in the number of distractor sources, whether distractors were speech or non-speech, the relative presentation levels of each source, their spatial configuration, and the dimensions of the room they were heard in. The cue was always an excerpt of the target voice in isolation, presented at the same location as the target in the corresponding mixture. Models classified the middle word within the target excerpt (out of 800 possible word classes).

Model architecture

Our main model architectures (referred to as "feature-gain models") incorporated learnable attentional gain functions into a deep neural network model of the auditory system (Fig. 1b). The neural network took simulated cochlear representations of an audio waveform as input. Feature-based attention was implemented with parameterized sigmoidal functions (green box in Fig. 1c) which mapped the model's internal representation of the cue to scalar multiplicative gains. These gains were then applied multiplicatively to the mixture representation.

The model architecture instantiated the hypothesis that attention is mediated by multiplicative gains that are deterministic functions of the features in a mental representation of the target of attention. To compute attentional gains, the cue stimulus for a trial was first passed through the auditory system model (blue model stages in Fig. 1b). The activations of each model stage were then time-averaged and stored to obtain stage-dependent memory representations of the cue (orange stages in Fig. 1b&c). Gains were computed per-stage as a sigmoidal function of the corresponding memory representations (features with high activations in the cue yield high gains), which were then multiplied with the activations of the mixture representation during the forward cascade of the model (Fig. 1b). Intuitively, features that are present in the cue should result in high gains, passing those features through the model when they occur in the mixture. The sigmoidal functions should enable models to learn how strongly to modulate the features at each model stage to select the target from the mixture. Gain function parameters and model features were jointly optimized, such that models should learn to prioritize information that enables selection of the target voice content.

We note that the application of feature gains happened immediately prior to each convolutional "block", which always began with a normalization operation. The model thus instantiated the key components of the normalization model of single-unit attentional effects[14], but embedded in a hierarchical system of learnable gains and features.

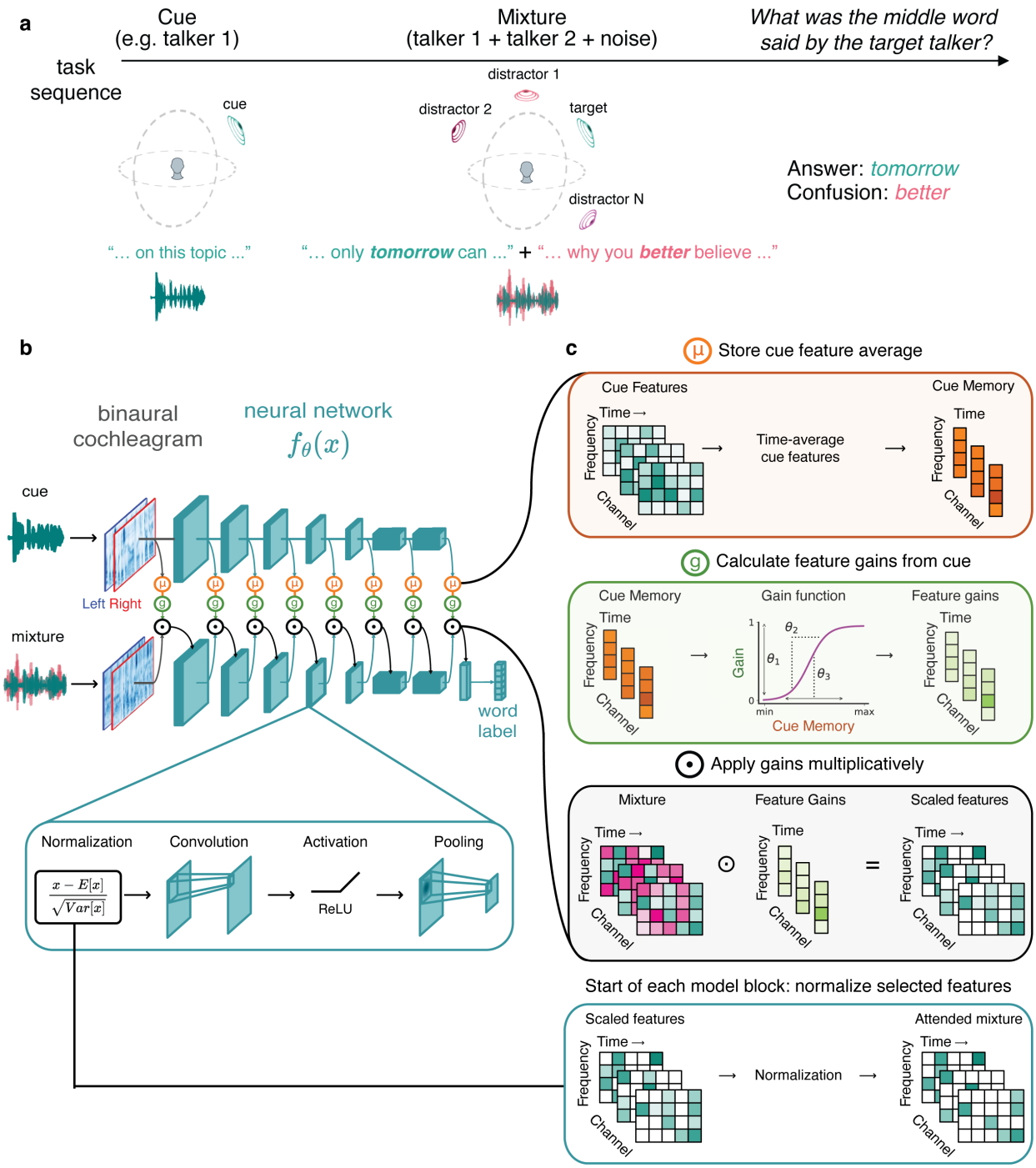


Figure 1. Task and model architecture used to study attention. a. Selective listening task. Listeners and models first heard an excerpt of a target talker's voice (the "cue"), then heard a mixture of a different excerpt of the target talker's speech superimposed with other sounds (the "mixture"), and then reported the word uttered by the target talker at the mid-point of the excerpt. b. Computational framework for modeling attention. A feedforward model of the auditory system (blue), comprised of a simulation of the cochlea followed by a neural network, took binaural audio as input and produced word labels as output. Each convolutional block of the model consisted of layer normalization, convolution, rectified linear activation, and pooling operations. The activations of the auditory model were multiplied by scalar gains (green). The gains were determined by sigmoidal functions that operated on the average activations of a cue stimulus (orange), obtained by passing the cue stimulus through the same model of the auditory system. c. Example application of multiplicative gains in model. Intuitively, the sigmoidal gain functions should enable gains to

be high for features that have high activations in the cue, allowing these features to be passed through the auditory system, enhancing the representation of the target talker.

To ensure the results generalized to some extent across the details of the model architecture, we trained 10 different feature-gain models with different architectures (Supplementary Table 1). Results for the feature-gain models are the average over these 10 architectures. In all models presented here we used a single gain function per model stage (we experimented with feature-specific gain functions, and found they did not improve performance). In a later section we test the extent to which the feature-gain architectural assumption was important to the results.

Human-model behavioral comparisons to assess attentional strategies

Because the models receive binaural audio signals, they should be able to learn to use both spatial cues and voice timbre cues if they are useful in solving the task. However, because the models were optimized only for word recognition, they were not explicitly constrained to reproduce human-like strategies, and were free to learn any solution that yielded good performance. To assess whether the model reproduced the properties of human attention, we conducted a set of experiments to characterize the model's behavior and to compare it to that of humans.

Model replicates human cocktail party performance in monaural conditions

We first ran human participants in a behavioral benchmark (Experiment 1) containing a series of diotic listening conditions (i.e., in which the same audio waveform was presented to the left and right ears, eliminating binaural spatial cues). We then simulated the same experiment on the model. Evaluation stimuli were new to the model (and to participants). As shown in Fig. 2, the model approximately replicated both the overall performance of human listeners and the dependence on signal-to-noise ratio and distractor type (see Supplementary Fig. 1 for results for the 10 individual model architectures).

Previous models of word recognition have replicated patterns of intelligibility across different types of noise[22,30]. However, because they lacked the ability to use a cue to selectively report one of several talkers, they were unable to perform at human levels in the presence of a single distractor talker, because without a cue, the task is ill-defined. Critically, the present model performed similarly to humans on trials including only the target and a single distractor talker (1-distractor condition in Fig. 2a). The similarity to human performance in this condition indicates that the model succeeds at attentional selection on par with human listeners.

Model replicates effects of distractor language and sex

To assess whether the model was sensitive to some of the same aspects of speech that influence human attentional selection, we reanalyzed data in the one-distractor condition (of Experiment 1). Human attentional selection is known to improve when the target and distractor talkers differ in sex, in part because of the acoustic differences this entails[26]. Selection is also better when distractor talkers speak a language unfamiliar to the listener[31,32]. Native English-speaking human participants in Experiment 1 replicated both of these effects, showing higher word recognition performance with different-sex distractors (Fig 2b) and Mandarin language (Fig 2c). The model also exhibited both effects, indicating the model learned to rely on some of the same cues as humans.

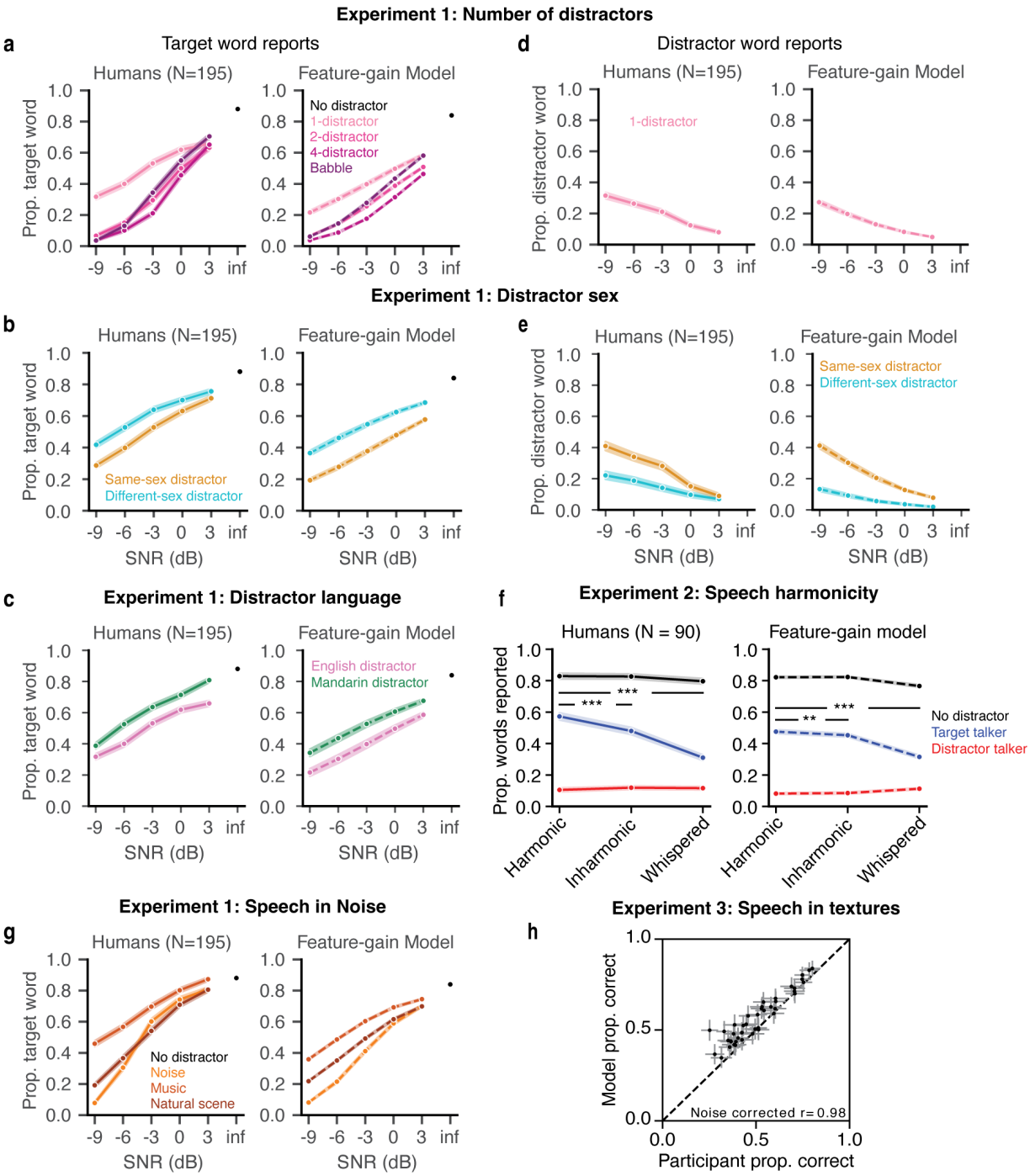


Figure 2. Comparison of human and model attentional selections in monaural conditions. a. Human and model performance vs. signal-to-noise ratio for speech distractor stimuli (Experiment 1). Here and in other panels, error bars plot SEM. Note that there is only one data point for the infinite signal-to-noise ratio condition, because the distractor is zeroed out in this condition, making the different distractor conditions equivalent. **b**. Effect of same- vs. different-sex distractor talkers (Experiment 1). **c**. Effect of foreign language distractor talkers (Experiment 1). **d**. Human and model confusions, plotted vs. signal-to-noise ratio (Experiment 1). Confusions occurred when the participant erroneously reported the word uttered by the distractor talker. **e**. Human and model confusions plotted separately for same- and different-sex distractor talkers (Experiment 1). **f**. Effect of talker harmonicity (Experiment 2). Note that the rate of confusions (reports of words uttered by the distractor talker) are low because the SNR was always 0 dB, at

which there are few confusions in Experiment 1 as well. **g**. Effect of non-speech distractor sounds (Experiment 1). **h**. Human and model performance across a large set of distractor noises (Experiment 3).

Model exhibits human-like failures of attention

Human attention is also known to demonstrate systematic failures. Even when attempting to attend to a target talker, humans sometimes mistakenly report what was said by a distractor talker instead. The root causes of these failures are not clear. In principle, they could represent lapses in attention, but they could also result from a distractor signal being sufficiently similar to a target signal that perfect selection is not possible.

To quantify selection failures, we re-analyzed the one-distractor condition of Experiment 1, measuring how often listeners reported words in the distractor utterance ("confusions" of the target and distractor). The overall confusion rate was low, but increased at lower SNRs ($F(4,776) = 51.707,\ p < 0.0001, \eta^2_{partial}$ = 0.210), and when the target and distractor were the same sex ($F(1,194) = 89.108,\ p < 0.0001,\ \eta^2_{partial}$=0.315; Fig. 2d&e). The model exhibited quantitatively similar effects (Fig. 2d&e). This result suggests that selection failures may be an inevitable consequence of target-distractor feature similarity.

Model replicates effects of speech harmonicity

Human attentional selection has also been shown to depend on whether the constituent frequency components of speech signals are harmonically related[33]. We ran an additional experiment to measure this effect using our task (Experiment 2), and simulated the same experiment on the model. Target, distractor, and cue signals were resynthesized to be harmonic, inharmonic, or whispered (always of the same type, e.g. all harmonic, all inharmonic, or all whispered). Without a concurrent distractor talker, human word recognition was similar for resynthesized harmonic, inharmonic, and whispered speech (Fig. 2f). But with a concurrent distractor, human recognition was modestly impaired for inharmonic speech, and substantially impaired for whispered speech, producing an interaction between the effect of speech harmonicity and single vs. mixtures of talkers (F(2, 178) = 51.021, p < 0.0001, $\eta^2_{partial}$=0.364), as in prior work (Fig. 2f). The model qualitatively reproduced these effects, also showing an interaction between speech harmonicity and single vs. mixtures of talkers (F(2, 18) = 116.151, p < 0.0001, $\eta^2_{partial}$=0.928), driven by worse performance for inharmonic ($t(9) = 5.752,\ p = 0.0003$, Cohen' s d = 1.819), and whispered ($t(9) = 15.142,\ p < 0.0001$, Cohen' s d = 4.788) speech compared to harmonic speech in concurrent distractor conditions.

Model replicates human speech-in-noise performance

The model also reproduced human-like patterns of performance across different types of noise distractors, like previous models[22,30]. In Experiment 1, both humans and the model demonstrated better performance with noise distractors compared to speech distractors, and for some types of noise compared to others (Fig. 2g). Fig. 2h shows an additional experiment (Experiment 3) comparing the model and human listeners across a large set of different types of (non-speech) noise[22], which further demonstrates that the model exhibits human-like dependence on noise type.

Models replicate signatures of human spatial attention

Humans also benefit from spatial separation between sources, an effect often termed "spatial release from masking". This benefit is thought to be partly attentional in origin, as it depends in part on knowing where to listen[27]. Because our model was trained on binaural audio rendered from sounds at different locations, it could in principle learn to use spatial information to aid

selection where helpful. However, because models were only optimized to recognize words, it was not obvious if spatial attention would emerge as a listening strategy.

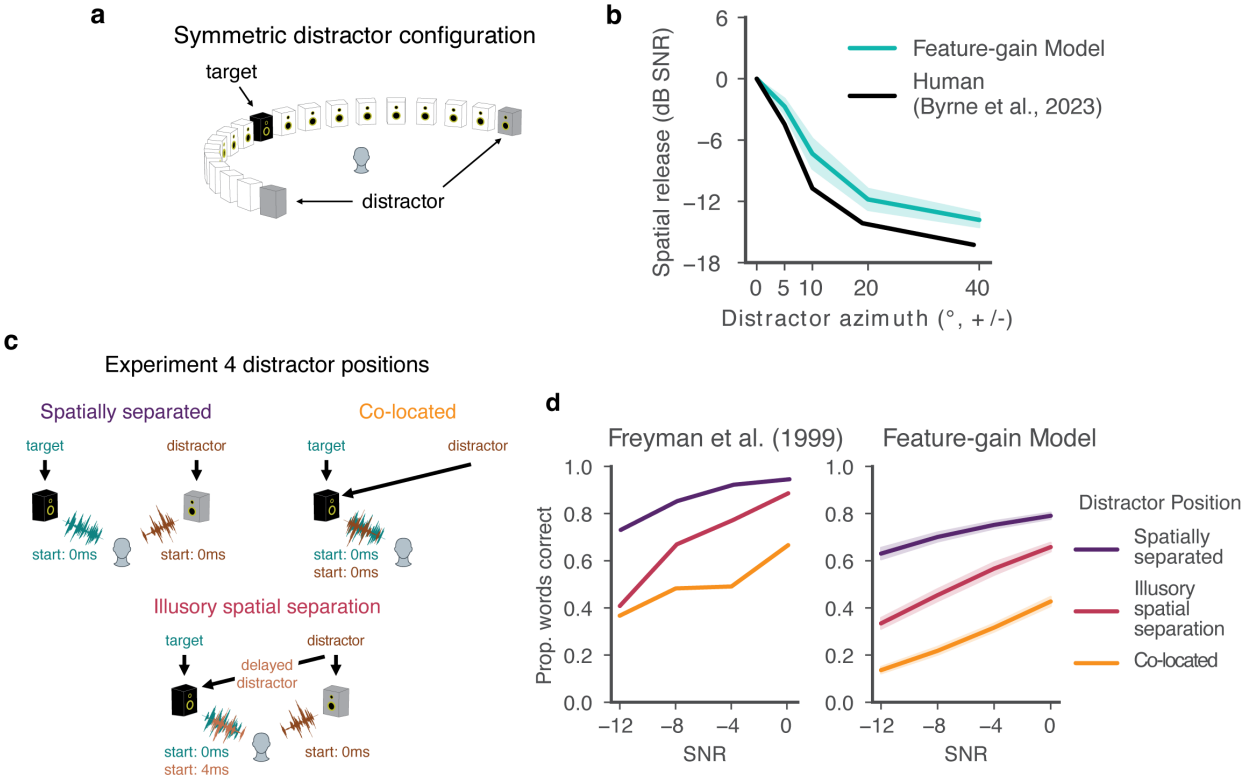


Figure 3. Comparison of human and model spatial attention. a. Stimulus setup for Experiment 4. The target talker was positioned in front of the listener, with distractor talkers positioned symmetrically on either side. Word recognition performance was measured as a function of signal-to-noise ratio for different target-distractor spatial offsets. **b**. Results of Experiment 4. Graph plots spatial release from masking, measured as the decrease in speech reception thresholds from the co-located condition where distractor talkers were also positioned directly in front of the listener. Human data were scanned in from original publication[34] and replotted. Error bars plot 2 SEM for the Feature-gain model (error bars for Humans not available in original publication). **c**. Stimulus setup for Experiment 5. Speech signals were played from two speakers. Target and distractor could be spatially separated, co-located, or given illusory spatial separation by virtue of the precedence effect. In this latter condition the target and distractor talker were both played from the left speaker, and a second copy of the distractor speaker was played from the right speaker. The co-located copy of the distractor was slightly delayed relative to the separated copy, inducing the illusion that the distractor came from the right speaker. **d**. Results of Experiment 5. Graph plots word recognition vs. signal-to-noise ratio for the three different spatial configurations. Error bars plot 2 SEM for the Feature-gain model (error bars for Humans not available in original publication).

To test if the models exhibited human-like effects of spatial separation, we evaluated performance as a function of target-distractor separation in azimuth. We replicated a previously published experiment in human listeners[34] in which performance was measured with distractors placed symmetrically in azimuth (Fig 3a). This design has the advantage of preserving the signal-to-noise ratio (SNR) at both ears as the distractors are moved away from the target talker, such that an advantage of spatial separation cannot be explained merely by increased SNR at one of the ears. As in the human experiment, we summarized performance at each spatial separation with a threshold (the signal-to-noise ratio granting 50% of ceiling performance). The model displayed thresholds that varied with spatial separation on par with humans (main effect of offset for model,

F(4,36) = 426.57, p<0.0001, $\eta^2_{partial}$=0.979; Fig. 3b), suggesting that it also made use of spatial location for selection.

Model replicates human advantage with illusory spatial separation

Another compelling demonstration of spatial attention in humans comes from the fact that human listeners benefit from illusory separation of sources. Such illusory separation can be mediated by the "precedence effect"[35,36], in which a sound played in rapid succession from two locations is heard to come from the first location, likely a consequence of a strategy for robust localization in the presence of reflections[37]. Specifically, a distractor signal co-located with a target signal (Fig. 3c, top right) impairs speech recognition relative to when it is spatially separated (Fig. 3c, top left). But if a second, spatially displaced copy of the distractor signal is played along with the co-located signal, human performance improves provided the displaced copy begins shortly before the co-located copy. The presumptive explanation is that the co-located copy is interpreted by the brain as a reflection, such that the distractor is represented as coming from the displaced location, and can be filtered out with spatial attention. To test whether the model similarly benefitted from illusory spatial separation, we simulated a previously published experiment[38] contrasting the benefit of illusory and true separation (Fig. 3c). As shown in Fig. 3d, model performance benefited from illusory separation in a qualitatively similar way as in humans (main effect of illusory separation vs. co-located, F(1,9) = 1224.917, p<0.0001, $\eta^2_{partial}$=0.993). As in models optimized for sound localization in realistic environments[37], the model presumably learns representations of sound location that are adapted to the presence of reflections, allowing it to benefit from the illusory spatial separation in this setting.

Model predictions of human behavior

There are many demonstrations of auditory spatial attention in humans, but they have mostly been restricted to a fairly consistent set of spatial configurations. One benefit of a stimulus-computable model is that experiments on the model are inexpensive, such that the model can be used to screen large numbers of experimental conditions to probe for interesting effects. Such effects can then be tested in humans, serving both to further characterize human perception, and as strong tests of the model. With these goals in mind, we ran the model on all possible combinations of target-distractor locations, summarizing the effect on performance in Fig. 4a (see Supplementary Fig. 2 and 3 for full set of results). Two effects stood out from this exhaustive set of tests, which we subsequently tested for in humans.

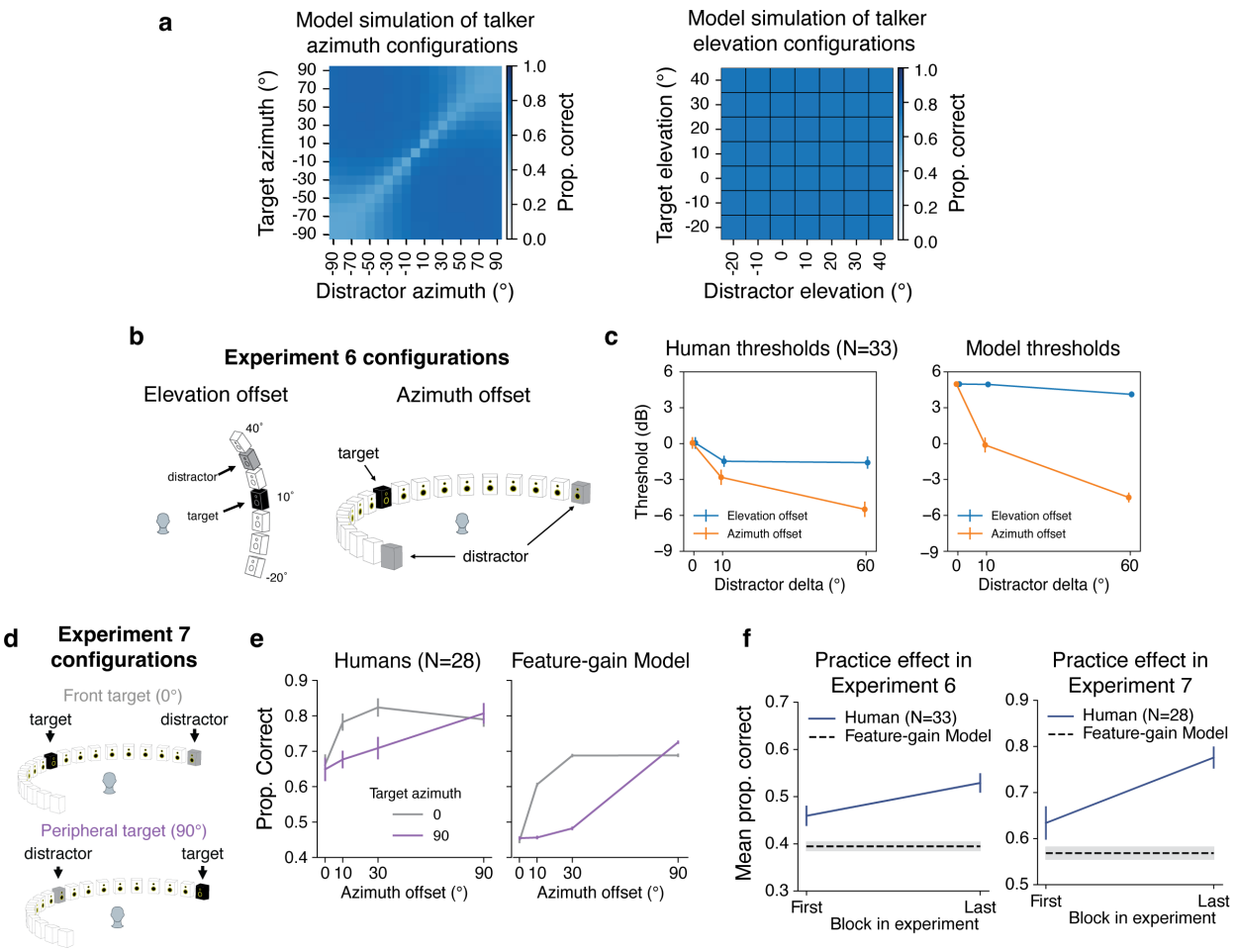


Figure 4. Model predictions of human spatial selection. a. Model word recognition performance across all possible combinations of target and distractor locations. Results are averaged over elevation and the front and back hemifields (left) or azimuth (right) for ease of visualization. Model results suggest larger benefits from azimuthal separation than elevation separation (note that there was almost no effect of separation in elevation, such that the right panel is uniform), and a "spotlight" that varies in width, being narrow for targets at the midline and wide for targets at peripheral locations. **b**. Spatial configurations tested in Experiment 6. To eliminate the possibility that an advantage from azimuth offset might be due to changes in signal-to-noise ratio in the input to an ear, the azimuth offset condition presented two distractor talkers, symmetrically located about the target talker position. Elevation offset condition presented two distractor talkers at one elevation (constrained by the locations that were possible given the speaker array). **c**. Results of Experiment 6. Humans and models show larger benefit of spatial separation in azimuth than in elevation. Here and in panels e and f, error bars plot SEM. **d**. Spatial configurations tested in Experiment 7. Target talkers were positioned at either 0 or 90 degrees, with distractors at one of four offsets. **e**. Results of Experiment 7. Error bars plot SEM. **f**. Practice effect evident in human performance in Experiments 6 and 7. Graphs plot human performance in the first and last block of each experiment, compared to overall model performance.

Model predicts horizontal/vertical asymmetry in spatial selection

When we exhaustively examined the effect of target-distractor configurations on the model, target word recognition increased as a function of the distractor spatial offset, as expected. However, offsets in the vertical direction produced much less masking release than the same offsets in the horizontal direction (Fig. 4a). Such differences are plausibly due to the different cues involved in the two types of offset – horizontal offsets are mostly signaled by differences between the ears, whereas vertical offsets are mostly signaled by monaural cues. It seemed plausible that monaural

cues might be less robust to the presence of concurrent sources. This issue has been addressed in a handful of previous studies[39-44], but with mixed results, and to our knowledge, the effect of horizontal and vertical offsets had never been compared in the same setting.

To test for analogous effects in humans, we ran an experiment with 33 participants measuring speech reception thresholds (i.e., the signal-to-noise ratio permitting average performance of 50% correct) for different target-distractor offsets in azimuth and elevation (Fig. 4b; Experiment 6). As expected, thresholds decreased as azimuthal distractor offsets increased, replicating known effects of spatial release from masking. However, the benefit was much weaker in elevation (Fig. 4c; interaction between direction of offset and extent of offset was statistically significant, $p < 0.0001$, using a non-parametric test for interaction, see Methods), confirming the model prediction.

Model predicts central/peripheral differences in width of spatial "spotlight"

A second effect that is evident in Fig. 4a is that the spatial separation between target and distractor needed to obtain a benefit is much less for targets at the midline (0°) than for targets in the periphery ($\pm$90°). See Supplementary Fig. 4 for further analysis of this effect for both speech and noise distractors. This effect is plausibly due to the higher acuity of localization at the midline[45-47], which in turn is thought to relate to the derivative of binaural cues with spatial position. However, it is not obvious that spatial acuity for isolated sources should directly translate to the "width" of spatial attention effects. To our knowledge, this issue had not been previously examined in humans.

We ran an experiment with 28 human participants measuring the effect of masking release as a function of distractor offset, for both centrally and peripherally positioned targets (Fig. 4d; Experiment 7). The model was run on a simulation of the same experiment to enable direct comparison. As shown in Fig. 4e, the qualitative difference between central and peripheral targets seen in models was also evident in humans: the distractor had to be spatially offset by a much larger amount for peripheral targets to yield the same selection benefit seen at smaller offsets centrally positioned targets, producing a significant interaction between the effect of target-distractor offset and target position ($F(3,81) = 6.731$, $p=0.0004$, $\eta^2_{partial}=0.200$).

We note that in both Experiments 6 and 7, human performance was overall somewhat higher than model performance. An analysis of human performance over the course of both experiments showed evidence of a practice effect in humans. Specifically, human performance was similar to model performance at the start of the experiment, but got better over time (Fig. 4f). A plausible explanation is that human participants were able to adapt to the room in which the experiment was conducted[48,49], boosting their performance (the model has no such ability as its weights were frozen following training, and did not change over the course of simulated experiments).

Models exhibit signatures of late selection

A signature of human auditory attention to speech is the enhancement of the neural representation of a target source at relatively late stages of the presumptive auditory hierarchy, such as non-primary auditory cortex[18,19,50-53]. However, it remains unclear why attentional enhancement might be limited to particular processing stages. In principle, the stage at which enhancement occurs could be determined by anatomical constraints on top-down connections. But they might also simply reflect the stage at which features are most useful for selection. Task-optimized models provide one way to gain insight into this issue, because they reveal an optimized solution that can be compared to that of the brain.

To assess the locus of attentional enhancement, at each model stage we measured correlations between the activations of target-distractor mixtures and the activations of either the target or distractor alone[16] (Fig. 5a). Attentional enhancement of the target should result in target-mixture correlations being higher than distractor-mixture correlations. Differences between target-mixture and distractor-mixture correlations emerged only at later model stages (Fig. 5b; see Supplementary Fig. 5 for results plotted separately for individual model architectures), indicating that enhancement occurs relatively late in the model. The same analysis performed on a model with randomly initialized weights did not show evidence for stage-specific enhancement (Fig. 5c), indicating that this result is not an inevitable consequence of the model architecture. This result is qualitatively consistent with human neuroscience evidence[18,19,51-53], and suggests that the solution arrived at in the brain is reflective of the representational locus at which enhancement is effective in enabling speech recognition.

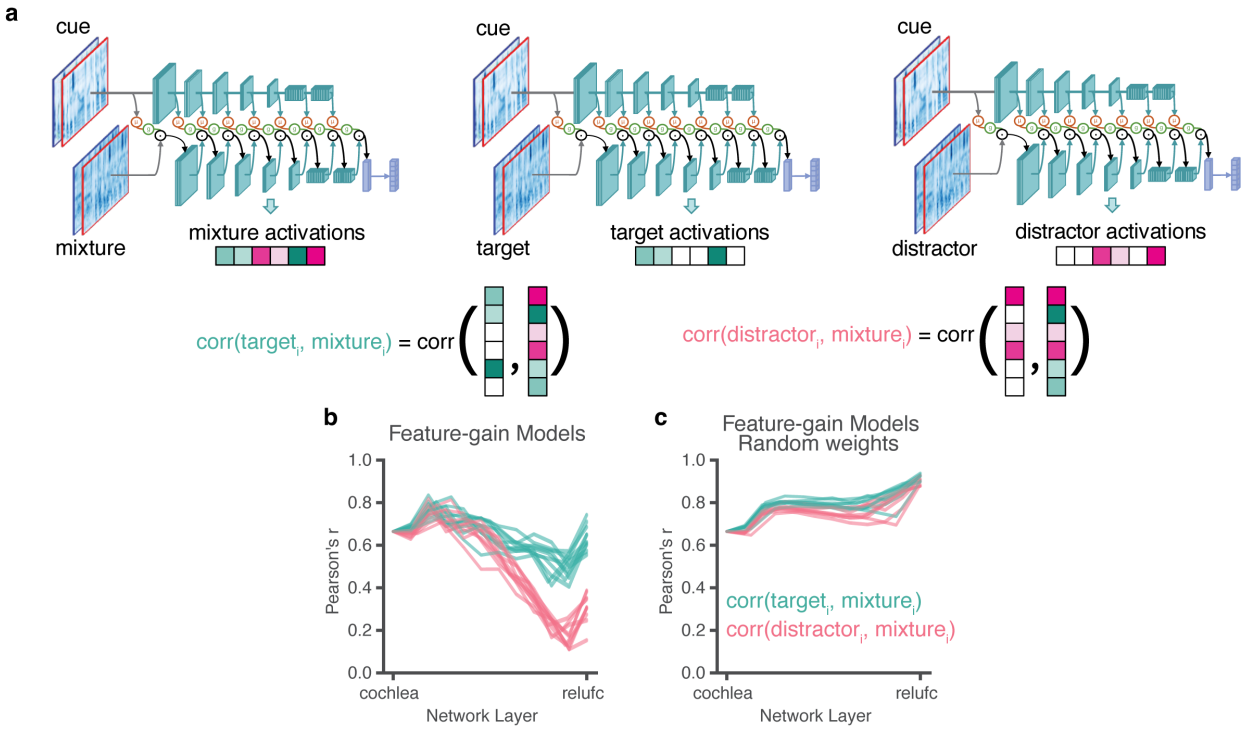


Figure 5. Stage of attentional selection. a. Explanation of stage-of-selection analysis. Target, distractor, and mixture excerpts were passed through the auditory model, with feature-gains derived from a cue matching the target. At each model stage, the activations of the target and distractor were correlated with the activations of the mixture. **b**. Task-optimized feature-gain models show signatures of target selection in late stages. Each line plots results for one of 10 model architectures. In all architectures, the late-stage representation of the mixture is more similar to that of the target than that of the distractor. In b and c, error bars on results for individual models are omitted for clarity. **c**. Same as b, but for models with random weights.

Models lacking architecturally constrained feature gains are less human-like

The inclusion of gain functions in the model instantiates an inductive bias on how attention could work. To investigate the importance of this inductive bias, we trained several alternate versions of the model (Fig. 6a). A “baseline” model removed the constraint of explicit gain functions entirely, instead receiving the mixture and cue as separate input channels, with model weights again optimized for recognition of the target word in the mixture. We also ran alternate architectures in which attention was constrained to operate only on particular stages. An “early-only” model had

gains applied only at the cochleagram stage, whereas a “late-only” model had gains applied only at the final fully connected model stage.

After optimizing each of these alternative models we ran them on each of the previously described experiments, measuring the proportion of target words correctly reported in each condition as well as the proportion of confusions in single-distractor conditions. We compared each model’s pattern of performance across conditions to that of humans, using both root-mean-squared error and correlation metrics. As an overall summary measure, we jointly analyzed both the proportion correct in each condition, and the proportion of confusions in each single-talker-distractor condition.

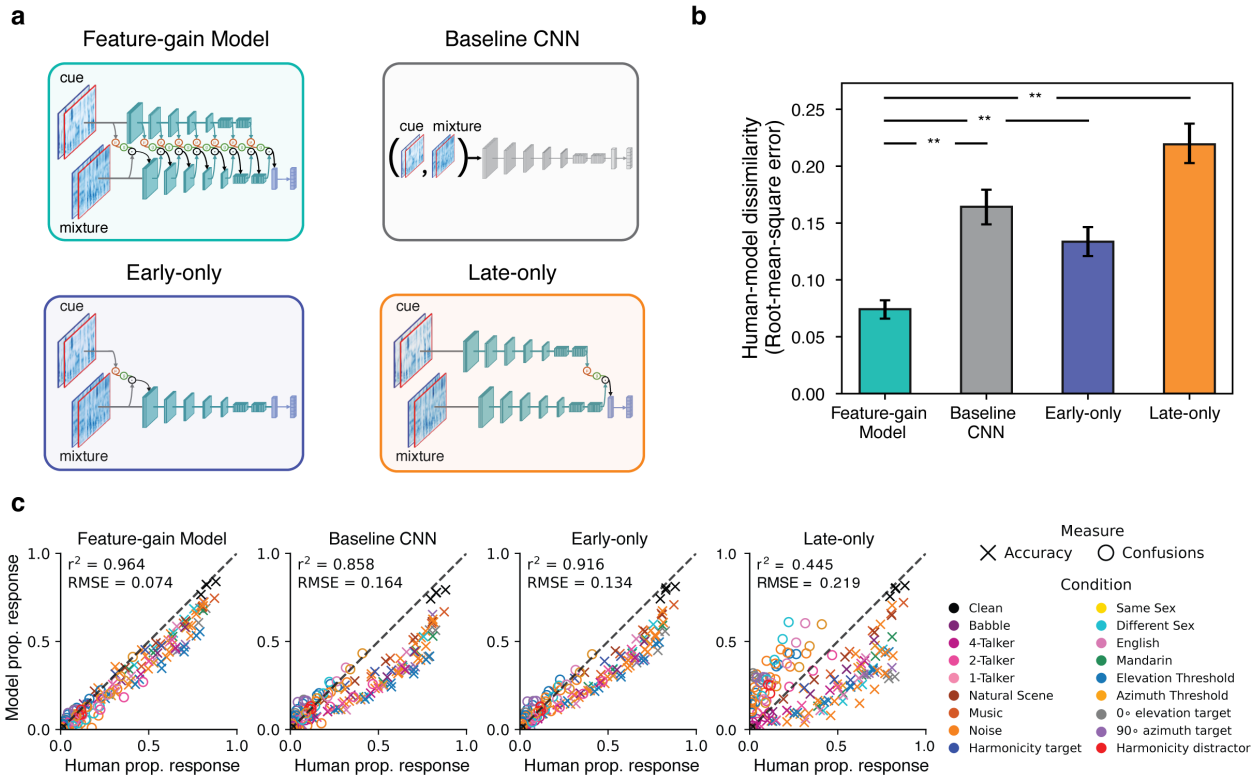


Figure 6. Dependence of attentional selection on model architecture. a. Alternative model architectures: i) the main feature-gain architecture; ii) a baseline architecture in which the cue was supplied to the model as an additional input channel, without explicit feature gains; iii) early-only architecture in which feature gains were applied only at the cochleagram; iv) late-only model in which feature gains were applied only at the last convolutional stage. **b**. Aggregate measure of human-model behavioral dissimilarity for each model architecture. Graph plots root-mean-squared error between human and model performance in all experimental conditions. Error bars plot 95th confidence intervals obtained via bootstrap of human-model similarity. **c**. Scatter plots of model vs. human performance in each of the experimental conditions from Figures 2-4. Each of the alternative architectures produces a worse match to human performance in some conditions.

Overall, the feature-gain model explained much of the variance in human performance across all experiments, and the explanatory power depended on the architectural constraints provided by the feature gains (Fig. 6b; see Supplementary Fig. 6 for results for each individual model architecture). Each alternate architecture showed significantly lower human-model similarity (root-mean-squared error: feature-gain vs baseline CNN p=0.002 difference of means = -0.070; feature-gain model vs early-only p=0.002, difference of means = -0.040; feature-gain model vs late-only p=0.002 difference of means = -0.125; Pearson’s r^2: feature-gain vs baseline CNN

p=0.002 difference of means = 0.093; feature-gain model vs early-only p=0.002, difference of means = 0.036; feature-gain model vs late-only p=0.002 difference of means = 0.507). Inspection of the pattern of proportion correct reveals that the alternative models tended to perform worse than the feature-gain model, but were also less correlated with humans across conditions. See Supplementary Fig. 7 for full results from Experiment 4, which reveals weaker spatial release from masking for each alternative model. These results confirm that the architectural bias of multiplicative gains helps to reproduce human-like attentional behavior.

Discussion

We investigated feature-based attention using stimulus-computable task-optimized models of the cocktail party problem. We augmented standard feedforward neural network architectures with memory-driven multiplicative feature gains, and optimized the models to report the word spoken by a cued talker given only binaural audio. The resulting models replicated the phenotype of human auditory attention, correctly reporting a cued talker's speech at comparable levels to humans, and exhibiting performance variation across conditions like that of humans. In particular, models exhibited human-like advantages for different-sex and different-language distractors, harmonic voices, and both real and illusory spatial offsets between target and distractor talkers. We used the models to sample target-distractor spatial arrangements more exhaustively than has been possible in human listeners, and saw two notable effects that we then tested in humans. Both of these model predictions were borne out in human listeners. The model also made errors in the same settings as humans, and to around the same degree. However, model performance, and human-model similarity, were dependent on the architectural motif of multiplicative gains, being worse in models without explicit gain functions, or with gains restricted to either early or late model stages alone. These results provide support for the idea that feature-based attention can be explained by multiplicative gains, and suggest that both human successes and failures of attention reflect an optimized solution to the problem of selecting a sound source via its features. Lastly, inspection of the model representations showed evidence of late selection, providing a normative explanation of effects seen neurophysiologically.

Relation to prior work

There has been little prior work incorporating attention into working models of sensory systems. Previous auditory models of the cocktail party problem have largely focused on the problem of inferring the distinct sources underlying an auditory scene, without a means to direct attention to one or more of the sources[54-60]. Previous computational work on attention has tended to either model effects of attention on neural responses[14,61] (rather than behavior), or has considered behavioral effects in small-scale models and simple tasks[62-67]. The main previous attempt to test the effect of multiplicative gains in a working model tested vision models on an object detection task in which four different images of objects were concatenated[68]. This study found that the application of gains proportional to a unit's selectivity for a particular object category increased the likelihood that the model would report that category, as might be expected to occur behaviorally in humans. A related recent vision model used feedback connections as a way to emphasize particular object categories, again showing that this aided a model's object detection in concatenations or superpositions of object images[69]. Our work is distinct in a) optimizing attentional mechanisms for task performance, b) testing the effect of attentional mechanisms in a naturalistic setting and task, c) showing that a single computational framework can account for many of the known attentional phenomena in the domain we consider (attention to speech), and d) using a model to make predictions about human attentional selection (and then validating these predictions).

Other work has used models to explore other aspects of attention. For example, neural network models can be trained to guide simulated eye movements during visual search using priority maps

computed from a target image, and reproduce aspects of human visual search behavior[70,71]. "Bottom-up" exogenous attentional cueing effects can also in some cases emerge naturally in task-optimized models[72]; such effects are widely believed to tap into distinct mechanisms from the endogenous effects we studied. We note that the word "attention" is also used to refer to a computational motif within transformer architectures popular in current machine learning[73], but that this motif differs from biological selective attention in not being directed to a particular target object or sound source.

Our work was inspired by a large body of neuroscience experiments documenting neural correlates of attentional selection[18,19,51-53,74] and suggesting that such correlates can be explained by multiplicative effects on neuronal responses[10,11]. However, such experimental findings leave it unclear whether the documented neurophysiological effects are sufficient to account for behavioral effects. Our results indicate that multiplicative gains applied at the appropriate stage of processing, with the right features, are sufficient to enable human-like auditory attentional behavior.

One of the main debates surrounding attention involves the stage at which attentional selection occurs[51,53,75-78]. These debates originally concerned whether attention acted before or after "semantic" processing, but with the discovery of attentional effects within sensory systems, interest shifted to the differences in attentional modulation between stages of sensory hierarchies[79]. Attentional effects are often more pronounced in later stages of sensory systems, but the reason for this has not been clear. Our results suggest that relatively late selection is an optimal solution for attentional selection, at least for the task of selecting a speech signal. It remains possible that the optimal stage of selection could depend on the task[80], which could be investigated by applying our framework to multiple tasks.

Our work builds on a large prior literature that has documented human performance in selective listening tasks with speech[25-29,31-33,38]. The model clarified these previously documented phenomena in several ways. First, the model results are consistent with the idea that human selective attention is near-optimal, in that the pattern of performance of a model optimized to select a target talker's speech was very similar to that of humans. In particular, the model made selection errors to about the same extent and in the same conditions as humans, indicating that such selection failures may be an inevitable consequence of feature overlap between target and distractor talkers. A priori it was not obvious that this would be the case. Second, we used the model as an engine to screen a large set of spatial configurations, not all of which had been tested thoroughly in humans. This process yielded predictions of effects that we then confirmed in human experiments. Such cycles of human and model experiments illustrate the value of having stimulus-computable models of behavior.

Limitations and future directions

At present our modeling framework requires a stereotyped task setting in which there is a cue stimulus from which attentional gains can be derived, followed by a stimulus to which the attentional gains are applied. This is an appealing setting in which to study attention, but it does not fully capture the variety of ways in which real-world attention arises. In many settings attentional gains seem likely to be refined over time, as the experience of attending to a particular sound source plausibly causes the representation of the source's properties to become more precise, which could be profitably incorporated into the attentional "filter". Evidence for an attentional filter that evolves over time comes from findings that humans can use attention to track sources whose features change over time[81,82]. Accounting for these abilities would require a more complicated way to set attentional gains based on the selected source. The ability to refine attention over time may be particularly important when attention is based on an abstract memory

of a class of target stimuli (e.g. the sound of a motor vehicle, or of a woman's voice), rather than a specific recently heard stimulus. The abstract memory may serve as an initialization of attentional gains, that are then updated based on the experience of listening to the actual sound source encountered by the listener.

Another limitation of our model is that attention is a deterministic function of the cue stimulus, based on a time-average of the cue features. In some settings, attentional gains might additionally be shaped by the properties of stimuli to be ignored. Attention is also likely subject to executive control. In particular, the strength of human attention can likely be varied by the observer, perhaps as is needed based on the perceived difficulty of a task. The strength of attention may relate to the feeling of effort. Extending the modeling framework to allow the strength of attentional gains to vary could help to understand effort in computational terms.

There is longstanding interest in the representation of stimuli outside the focus of attention[1], but thus far models have not been available to provide hypotheses for what should be expected to be represented given candidate computational instantiations of attention. A natural next step would be to measure what can be decoded about unattended stimuli from different model stages, validate such measurements against existing data[17], and use the models to make new predictions that can subsequently be tested against brain data.

The models we built here are composed of simple operations that are loosely inspired by neuroscience, but deviate in many ways from biological sensory systems, making them inappropriate as models of some neural phenomena related to attention[83]. However, as it becomes possible to train models that are more biologically realistic[84], the general framework we propose here should remain applicable. Lastly, our framework is modality-independent, and extensions to other modalities could help reveal whether general principles govern feature-based attention in vision and hearing.

Methods

Training data generation

The training dataset consisted of 3,973,192 labeled exemplars, with each exemplar consisting of a cue signal and a multi-source mixture with a corresponding word label. Cue signals were two-second binaural audio clips of a single "target" talker spatialized to a location (defined by an azimuth and elevation) relative to a simulated listener position. Mixture signals were another two-second binaural audio clip: a superposition of a different excerpt of the target talker (rendered at the cued location) and other spatialized audio signals (either excerpts of other talkers and/or non-speech sound sources, rendered at locations that could be distinct from that of the target). Word labels corresponded to the word spoken by the target talker that overlapped the middle (i.e., the 1-second mark) of the mixture. Below, we will describe the curation of source materials included in training, the room simulator used to spatialize sources, and the scene generation procedure.

Speech corpora

Speech excerpts used to train models were sourced from the English-language training split of the 9th release of the Common Voice speech corpora. This corpus was selected for its large number of unique talkers and variety of speech materials. We screened the corpus to obtain a curated list of speech excerpts that would support our attentional word recognition task. The purpose of the screening was to balance occurrence of single-word classes, the number of utterances per talker, and talker sex for both targets and distractors across examples contained in the training set.

First, word boundaries were extracted from the recordings and transcripts using a forced alignment procedure described in previous publications[85,86]. We then removed examples that did not have a 44,100 kHz sampling rate, helping to ensure that excerpts natively contained the frequency content implicated in supporting human sound localization[87]. Words spelled with less than five characters or spanning longer than two seconds in duration were excluded. From the remaining set of word excerpts, we took at most 5,000 examples per word to limit word class imbalance, randomly sampling from the available examples for word classes exceeding this limit. The top 800 most numerous words were taken to be our training vocabulary. Word class balance was then obtained by resampling each word class to have 5,000 exemplars, drawing from the screened set with replacement. This returned 3,994,484 million total single-word examples, with 2,041,825 unique utterances before up-sampling word classes.

Of the 15,735 unique talkers that remained in our training set after the prior screening steps, 3,407 were female talkers and 12,325 were male talkers. To avoid overrepresenting a particular talker sex when generating training scenes, our final screening step was to filter the 3,994,484 million excerpts to obtain a sex balance at the excerpt level. We did this by taking all available examples of female talkers in the word-balanced set (496,649 of 3,994,484), then sampling the same number of male-talker examples from the remaining set. The resulting screened set holds 993,298 total single-word examples (496,649 female, 496,649 male). Cue, target, and distractor speech excerpts were sampled from this screened set of single-word examples when constructing the final training set of cocktail party scenes.

Validation set examples were generated using the same screening procedure described above, sourcing materials from the corresponding validation split of Common Voice. For the validation set, we sampled up to 250 examples per word class instead of 5,000 as in training, resulting in 196,968 single-word excerpts. Gender balancing was not imposed for the validation set.

Noise corpora

Natural sounds clips were sourced from AudioSet[88]. We screened the entirety of AudioSet to select sound excerpts that were appropriate for our scene generation procedure. The main goal was a) to avoid speech content, as we wanted all speech in the dataset to come from the speech corpora described above, and b) to ensure that the sampling rate was high enough to be appropriate for spatialization. We first screened AudioSet examples to find a curated list of "parent" clips, then excerpted individual clips from these parent clips. We first removed AudioSet examples labeled with "Music", "Speech", "Singing", "Vocal music", "Whispering", "Shout", or "Silence". As with the Common Voice excerpts, we then removed examples that did not have a 44,100 kHz native sampling rate. Parent clips were also screened to be at least 9-seconds in duration, enabling many 2.5-second excerpts per example. We performed the screening separately on the original training and validation split assignments of AudioSet. This returned 661,877 parent clips for training and 14,570 parent clips for validation. Individual natural sound clips were taken from these parent clips as part of the scene generation procedure. To obtain each natural sound clip in a scene, we first sampled a parent clip (uniformly from the screen training clips), then sampled a 2.5-second excerpt from the parent clip via a uniform random crop.

Virtual room simulation

To spatialize scenes, we used the same room simulator and a similar set of room simulations as Saddler and McDermott (2024)[22]. The description of the simulator is reproduced from their paper apart from minor edits indicating where our parameters differ from theirs. The simulator used the image-source method, incorporating KEMAR HRTFs, to render sets of binaural room impulse responses (BRIRs) for 2000 different shoebox-style rooms. Room dimensions were sampled log-uniformly between 3 and 30 m for length and width, and between 2.2 and 10 m for height. The listener's head position was sampled uniformly in each room, under the constraint that the head was at least 1.45 m from every wall and no higher than 2 m from the floor. BRIRs for 1584 source locations (2 distances from the listener × 72 azimuths × 11 elevations) were rendered for each listener environment. One of the distances was 1.4 m for every BRIR. The other distance was independently sampled for each BRIR (drawn uniformly between 1 m and 0.1 m 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. The final training and validation datasets consisted of 2,851,200 and 316,800 simulated positions, respectively. Our simulated spatialization differed from that in Saddler and McDermott (2024) only by the inclusion of negative elevations: we used 11 total elevations compared to the 7 in Saddler and McDermott, giving 1,584 source locations per listener environment with 2,851,000 total positions (vs. 1,008 source locations per listener environment and 1,814,400 total positions in Saddler and McDermott).

Training data scene generation

Auditory scenes were created by combining talker clips from the curated set of Common Voice examples with natural sound clips from the curated set of AudioSet examples, which were then spatialized using the room simulator. Scene parameters were varied across training exemplars to sample a wide range of conditions. To generate a training exemplar, we used the following procedure. First, we sampled a target excerpt from the curated list of Common Voice examples. Second, a cue excerpt was sampled from the set of clips produced by the target talker, restricted to be centered on a different word than the target clip. Third, we sampled the total number of distractor sources (uniformly from 1 to 6 sources). Fourth, we sampled the number of these distractor sources that were talkers (uniformly from 0 to n, with n being the sampled number of distractors for the scene). Fifth, we sampled the distractor talker clips from the Common Voice examples not produced by the target talker. Sixth, we sampled the remaining distractor sources from the curated set of AudioSet excerpts. Individual natural sound clips were taken from an AudioSet excerpt by first uniformly sampling a parent clip, then randomly excerpting a 2.5-second

crop from that parent clip. Seventh, the sound pressure level of each distractor source was uniformly sampled from 50 dB SPL to 70dB SPL.

In each scene, the target and cue source were localized to the same azimuth and elevation, relative to the sampled listener position. To prevent the model from exclusively exploiting localization cues, half of the training scenes had distractor sources at the same location as the target source. For the remaining half of training examples, the azimuth and elevation of each distractor source were uniformly sampled from the possible source locations relative to the listener position. After setting distractor clip levels, distractor clips were spatialized to their selected locations then combined (summed in each channel).

To increase variability in the training data, each target clip in the curated set of Common Voice was sampled four times, each time being part of a different scene. The final composition of the training set contained 3,973,192 unique exemplars (1,988,317 female; 1,984,875 male). Validation scenes were generated via the same procedure, but with clips sampled from the validation set source materials, resulting in 196,968 exemplars.

Signal augmentations

Cue, target, and distractor-scene compositions were pre-sampled and stored separately, enabling augmentations to be applied to the target clips and distractor scenes. To prevent models from conflating the characteristics of the recording conditions for an individual talker with properties of their voice, bandpass filters were applied as augmentations to either the cue or target excerpt for 50% of the training examples. Filtering was performed using digital Butterworth filters, stochastically choosing filter parameters for each example, uniformly sampling low-frequency cutoffs from 40Hz to 400Hz, high-frequency cutoffs from 4kHz to 16kHz, and the filter order from 1 to 4. To avoid having models overfit to the onset times of labeled words, target clips were randomly shifted in time (either forward or backward with equal likelihood). Time shifts were uniformly sampled from 0 to 50% of the labeled word's duration, constrained so the labeled word still overlapped the middle (1-second mark) of the target clip after shifting. Mixture clips were obtained online during training by superimposing the target and distractor scenes at a signal-to-noise ratio (SNR) uniformly sampled from -10 dB SNR to 10 dB SNR. Cue and mixture clips were then RMS (root-mean-square) normalized to 0.02. Finally, to allow models to learn how to report words said in isolated speech excerpts without being cued, 10% of training examples features single-talker target signals with silence as a cue. This was done by simply not mixing the target with the paired distractor and using silence (an array of zeros) instead of the paired cue signal for these examples.

Boundary handling

To avoid signal onset/offset artifacts, all cue, target, and mixture signals were extracted to be 2.5 seconds in duration. Once augmentations were applied, signals were passed through the cochlear model, after which the middle 2-seconds were excerpted as the final stimulus.

Model implementation

Models were implemented in the PyTorch deep learning library using the PyTorch Lightning framework for efficient training in distributed settings. All model training and analysis used the compute resources of MIT's OpenMind computing cluster, running Python version 3.12, Pytorch version 2.4, and Pytorch Lighting version 2.1. All Python dependencies and package versions are available in the code repository that will be made available upon publication.

Cochlear model stage
The first stage of our models was a fixed simulation of the cochlea and auditory nerve, providing a ear-by-time-by-frequency representation intended to replicate the auditory cues provided by the human auditory periphery. This initial stage took the sound waveform as input. The input was passed through a finite-impulse-response (FIR) approximation of a gammatone filter bank (with impulse responses truncated to 25 ms to reduce memory consumption) the output of which was half-wave rectified, passed through a compressive nonlinearity, then downsampled. These operations were performed separately on the left and right audio channels.

First, each of the two 44.1kHz stereo audio waveforms was convolved in the time domain with the FIR approximation of a 40-channel gammatone filter bank (1,102 taps per filter) with center frequencies spaced uniformly on an ERB-numbered scale between 40Hz to 20kHz. Second, the resulting subbands were half-wave rectified. Third, the half-wave rectified subbands were raised to the power of 0.3 to simulate the compressive response mediated by outer hair cells. Fourth, the compressed, rectified subbands were low-pass filtered with a 4kHz cutoff and down-sampled to 10kHz, to both impose the upper limit on phase locking of inner hair-cells and reduce the dimensionality of the neural network inputs. Low-pass filtering and down sampling was performed via 1d convolution with a Kaiser-windowed sinc filter, with a filter width of 64, roll-off of 0.94759, and beta of 14.76965, implemented with the Torchaudio transforms resample method. Finally, to avoid signal onset/offset artifacts, the middle 2-seconds were then excerpted from the full 2.5-second input signal duration. We refer to the resulting representation as a "cochleagram". The left and right cochleagrams were concatenated, yielding a 2-channel by 40-frequency by 20,000-timestep input to the neural networks.

Neural network model stages
Models of feature-based attention were built using a convolutional neural network backbone equipped with feature-based gains as parameterized sigmoid functions. The neural network backbone consisted of a cascade of convolution, pooling, and normalization operations, as in previous models from our lab[22,30,37,89,90] which have yielded strong task performance and close matches to human behavior on other auditory tasks, as well as state-of-the-art predictions of brain responses from human auditory cortex[30,91]. Below, we will detail the core operations included in our artificial neural network architectures, the implementation of feature-based gain, the forward-pass used in our training algorithm, the architecture search process used to select model architectures, and the modifications made to constrain attention in our control models.

Feature-based gain implementation
Feature-based attention was implemented as sigmoidal functions with learnable parameters at each neural network model stage:

$$g^{(l)} = \sigma\left(m_{cue}^{(l)};\ \boldsymbol{\theta}_g^{(l)}\right) = \theta_1 - \frac{(1-\theta_1)}{1+e^{-\theta_2\left(m_{cue}^{(l)}-\theta_3\right)}},$$

where (l) indicates the network stage, $\boldsymbol{\theta}_g = [\theta_1, \theta_2, \theta_3]$ are learned parameters for the bias, slope, and threshold of the function, $m_{cue}^{(l)} \in \mathbb{R}^{C \times F}$ is the time-averaged representation of the cue at layer (l) (of size C channels by F frequencies), and $g^{(l)} \in \mathbb{R}^{C \times F}$ are the feature-gains for each of the $C \times F$ features. Gain functions were applied to the cochleagram and to each convolutional block in the neural network.

Forward pass through the feature-gain models
The forward pass for our feature-gain models augmented the standard forward pass used in feed-forward networks, enabling stage-specific application of gains derived from the cue representation

at a given stage to the mixture representations at the same stage. Given a cue x_{cue} and mixture x_{mix}, the forward pass of the model runs as follows. First, we obtain a representation of the cue for each model stage

$$x_{cue}^{(l)} = f^{(l)}\left(x_{cue}^{(l-1)};\ \boldsymbol{\theta}_a^{(l)}\right),$$

where $f^{(l)}\left(\cdot;\ \boldsymbol{\theta}_a^{(l)}\right)$ is the l_{th} model stage (i.e., convolutional block), $\boldsymbol{\theta}_a^{(l)}$ are the parameters for that model stage, and $x_{cue}^{(l-1)}$ is the cue representation from the previous model stage. Second, a memory representation, $m_{cue}^{(l)} \in \mathbb{R}^{C \times F}$, of the cue, $x_{cue}^{(l)} \in \mathbb{R}^{C \times F \times T}$, is obtained by averaging over the time dimension of the cue representation at the same model stage:

$$m_{cue}^{(l)} = \frac{1}{T}\sum_{t=1}^{T} x_{cue}^{(l)^{(t)}},$$

Third, gains are then obtained from the memory representation

$$g^{(l)} = \sigma\left(m_{cue}^{(l)};\ \boldsymbol{\theta}_g^{(l)}\right)$$

where $\boldsymbol{\theta}_g^{(l)}$ are parameters for the gain function at stage l. Fourth, the mixture representation is obtained

$$x_{mixture}^{(l)} = f^{(l)}\left(x_{mix}^{(l-1)};\ \boldsymbol{\theta}_a^{(l)}\right) \odot g^{(l)},$$

where $x_{mix}^{(l-1)}$ is the mixture representation from the previous stage, $f^{(l)}\left(\cdot;\ \boldsymbol{\theta}_a^{(l)}\right)$ is output of the l_{th} convolutional block, $g^{(l)}$ are the feature-gains obtained from the cue, and $\odot$ is the element-wise multiplication operator. For simplicity, the same set of feature-gains were applied to all timepoints of the corresponding mixture representation. After the final convolutional stage, the mixture representation was passed through a fully-connected layer followed by the 800-dimension linear output stage for the word recognition task.

During this forward pass, the weights of the convolutional blocks, $\boldsymbol{\theta}_a$, are shared when obtaining cue and mixture representations for each training example.

Artificial neural network architectures

The operations in the convolutional neural network architectures were organized as a series of convolutional blocks. Each convolutional block consisted of a sequence of layer-normalization, convolution, point-wise nonlinearity, and pooling operations. The final convolutional block of each architecture was followed by a sequence of four operations: a single fully-connected layer, a point-wise nonlinearity, dropout, and a softmax classifier.

Layer normalization

Normalizing activations between neural network stages improves the efficiency of ANN training by helping to stabilize gradient updates: the magnitude of a parameter update in one stage is less likely to be amplified in following stages if they are separately normalized. Normalization-like operations are also common in biological sensory systems, and have been proposed to interact with multiplicative gains to produce attentional effects observed neurophysiologically[14]. Layer normalization[92], a common choice for normalization operations in ANNs, point-wise normalizes input examples individually using the mean and variance over feature dimensions, with high output reflecting high activity relative to the features of that input example. We used layer normalization (as opposed to batch normalization[93], the other common choice for ANN normalization operations) because it is more similar to the normalization found in sensory systems in normalizing responses by a function of the current stimulus[94], rather than by a function of the training distribution. Layer normalization also does not require scaling test examples to match the

statistics of the training examples (as is necessary with batch normalization), which might aid generalization at inference. The layer normalization operation is defined as

$$x_{\text{normalized}} = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta$$

where $x \in \mathbb{R}^{C \times F \times T}$ is the input tensor, $E[x]$ and $Var[x]$ are the mean and variance over feature dimensions (C, F, T) of x, $\epsilon = 0.00001$ to prevent division by zero, and $\gamma \in \mathbb{R}^{C \times F \times T}$ and $\beta \in \mathbb{R}^{C \times F \times T}$ are tensors of learnable parameters.

Convolution

Each convolutional layer consisted of a bank of learnable filter kernels $W \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times n_f \times n_t}$, with C_{out} different kernels, C_{in} input channels, and kernel dimensions of n_{f} by n_{t} in frequency and time, respectively. Inputs to each convolutional layer, $x \in \mathbb{R}^{C_{\text{in}} \times F \times T}$, were three-dimensional tensors with C_{in} input channels, F features, and T time samples. For the first convolutional layer, $F = 40$ and $T = 20{,}000$ (the frequency and time dimensions of the cochleagram), while $C_{\text{in}} = 2$ (the left and right audio channels).

Boundary handling for convolution operations was identical to that in Francl and McDermott (2022). Specifically, "valid" convolution was used in the time dimension (i.e. no zero-padding applied), and "same" convolution was used in the frequency dimension. The rationale for this choice was to avoid temporal edge artifacts that would otherwise result from zero-padding in the time dimension. Edge effects in the frequency dimension are less clearly inconsistent with biology because the cochlea has upper and lower frequency limits, and were considered preferable to the rapid loss of dimensionality along the frequency axis that would occur with "valid" convolution given the small number of input frequency channels.

For an input tensor x, the output of a convolutional layer is a tensor $y \in \mathbb{R}^{C_{\text{out}} \times F \times (T - n_{\text{t}} + 1)}$ given by:

$$y_{k_{\text{out}},i,j} = \sum_{n=1,\ m=1,\ C_{\text{in}}=1}^{n_{\text{f}},\ n_{\text{t}},\ C_{\text{in}}} W_{C_{\text{out}},C_{\text{in}},n,m} \cdot x_{C_{\text{in}},i+n,j+m}$$

The output from each convolution thus had dimension $(C_{\text{out}} \times F \times T - n_{\text{t}} + 1)$. Because convolutional layers were preceded by layer-normalization, bias vectors were omitted (to minimize redundant parameters and reduce memory consumption). All convolutional layers used a stride length of 1.

Point-wise nonlinearity

Nonlinear activation functions enable neural networks to learn complex functions. We used rectified linear units as the point-wise nonlinearity in all architectures. This operation is defined as

$$ReLU(x) = \max(0, x)$$

Weighted-average pooling

Pooling layers downsample their inputs by aggregating information across neighboring frequency and time points. Weighted average pooling with Hanning windows was used to reduce aliasing in our networks (which would occur if downsampling was not preceded by lowpass-filtering)[95]. Pooling was performed by convolving input tensors $x \in \mathbb{R}^{C \times F \times T}$ with two-dimensional (frequency by time) Hanning window kernels, $H \in \mathbb{R}^{h_f \times h_t}$:

$$y_{\mathrm{k}} = H_{s_f, s_t} * x_{\mathrm{k}}$$

where $*$ is the convolution operation, k indexes the channel dimension, and s_f, s_t are the stride length in frequency and time, respectively. As in Saddler et al., (2021)[89], the Hanning kernel had a stride-dependent shape, where

$$h_{\cdot} = \begin{cases} 1 & s_{\cdot} = 1 \\ 4 \cdot s_{\cdot} & s_{\cdot} > 1 \end{cases}$$

for h_f and h_t depending on s_f and s_t, respectively. For an input $x \in \mathbb{R}^{C \times F \times T}$, the corresponding output is $y \in \mathbb{R}^{C \times \frac{F}{s_f} \times \frac{T}{s_t}}$. If either s_f or s_t equal 1, no pooling is performed on the corresponding dimension.

Fully-connected layer

A fully-connected (also sometimes called linear or dense) layer applies an affine transformation to an input tensor $x \in \mathbb{R}^D$. In our networks, where $x \in \mathbb{R}^{C \times F \times T}$ is the output of a convolutional block, x is first flattened to a vector $x_{\text{flat}} \in \mathbb{R}^D$ where $D = C \cdot F \cdot T$. Then, an affine transform is applied

$$y_i = \sum_{d=1}^{D} W_{i,d} \cdot x_d + b_i$$

producing an output vector $y \in \mathbb{R}^N$, where $W \in \mathbb{R}^{N \times D}$ are learned weights, and $b \in \mathbb{R}^N$is a learned bias vector.

Dropout regularization

Dropout is a form of regularization applied to input tensors $x \in \mathbb{R}^N$ during training, intended to minimize co-adaptation of units. On each forward pass, a fraction p of the units in x are sampled (from a Bernoulli distribution) and set to zero. The remaining fraction of units are scaled by $1/(1-p)$ so the expected sum over all outputs is the same as the expected sum over x. During inference, the operation is replaced by the identity function. Dropout was applied during training to the activations of the penultimate fully-connected layer preceding the softmax classifier in our networks, with $p = 0.5$.

Softmax classifier

A softmax classifier was the final layer in all our networks. The softmax classifier is a composition of a fully-connected layer and a normalized exponential function. First, input tensors $x \in \mathbb{R}^D$ are passed through a fully-connected layer $h(\cdot, W)$, with weights $W \in \mathbb{R}^{V \times D}$ mapping from D features to V word classes. The operation returns un-normalized activations for each word class (often called logits):

$$x_{\text{logits}} = h(x, W)$$

These logits are then scaled by the softmax function producing an output vector $y \in \mathbb{R}^V$:

$$y^{(v)} = \frac{\exp\left(x_{logits}^{(v)}\right)}{\sum_{i=1}^{V} \exp\left(x_{logits}^{(i)}\right)}$$

where v indexes the word class. Because the values of y are greater than zero and sum to one, y can be interpreted as a probability distribution over word classes for the given input mixture.

Architecture selection

Artificial neural network performance depends both on the network weights learned during gradient descent, and the configuration of the operations (defined by "hyperparameters") that define an architecture (i.e. the number of layers, number of channels per layer, and the operations in each layer). To ensure we used optimal hyperparameter settings in our CNN backbone, we drew 13 architectures from successful models of word recognition and sound localization identified by prior work[22,30,37,89,90]. First, we piloted all training and model experiments using one of these architectures (the top architecture in Saddler & McDermott, 2024[22]), which yielded strong matches to human behavior. Then, we trained the remaining 12 and selected the top 10 architectures based on their validation set performance. Model results for each experiment are reported as the average across these 10 best network architectures, enabling us to report measures of uncertainty and marginalize across the eccentricities of any single network architecture. Hyperparameters of these 10 architectures are in Supplementary Table 1.

Models with alternative architectural constraints

To test whether explicit feature gains were necessary for attentional selection, we instantiated models that altered how gains were included in the model architecture. We tested three control architectures that each informed a particular hypothesis. The first was a "gain-free" architecture, that tested the necessity of including explicit gain functions. The second was a model with gains only at the early stage of the architecture, that tested whether selection of low-level features was sufficient for task performance. The third control model included gains only at the late stage of the architecture, testing whether selection of higher-level features was sufficient for task performance. Each control architecture used the backbone CNN of the best-performing feature-gain model (architecture 1 in Table 1). Control models were trained using the same training set and optimization hyper-parameters as the feature-gain architectures. The details of each control architecture are given below.

Baseline model without explicit feature-based gains: We trained a version of architecture 1 using the backbone CNN without gain functions. Instead, cue and mixture signals were concatenated along the channel dimension and passed as a single input $x \in \mathbb{R}^{C\ x\ F\ x\ T}$with dimensions $C = 4$ channels (the stereo channels for cue and mixture), $F = 40$, and $T = 20{,}000$ (the frequency and time dimensions of each cochleagram). The model architecture was therefore augmented to accept 4 input channels (compared to 2 as in the feature-gain models), using the forward pass of a traditional CNN. This model tests the extent parameters of the CNN architecture alone support solutions to the attentional selection task, providing a performance baseline models with feature gains can be compared to.

Model with feature-based gains only at early stages: Inputs to the model, the model forward pass, and application of gains were identical to the unaltered feature-gain models, with the exception that gains were only applied to the cochlear representation of the mixture before the first convolutional layer. This model tests the hypothesis that constraining selection to low-level features available in the early auditory pathway is sufficient for enabling attentional selection.

Model with feature-based gains only at late stages: We modified architecture 1 to include attentional gains only between the last convolutional layer and the fully-connected layer. Model inputs, the model forward pass, and application of gains were identical to the full feature-gain models, with the exception that gains were only applied to the outputs of the final convolutional

layer. Consequentially, this model tests whether selection of features from higher-order representations alone is sufficient for enabling attentional selection.

Model optimization

All models described above were trained using the dataset described above via stochastic gradient descent, using the AdamW[96] optimizer (with a learning rate of 0.00005 and a batch size of 288). Models were trained using a distributed data parallel strategy run on 4 NVIDIA A100 GPUs, where each GPU ran a unique subset of a training batch through a copy of the model parameters (which were updated synchronously with respect to the whole-batch loss) enabling larger batch sizes in training. 16 CPUs (4 per GPU) and a total of 100Gb of memory were used to execute data reading and online signal augmentations during training for each model. All models were trained until performance on the validation set converged (approximately 10 training epochs of the training set).

Human behavioral experiments

Human experiments were conducted both online and in-person, depending on the type of experiment. All participants in both online and in-person experiments self-reported as native English speakers having no known hearing loss.

All diotic listening experiments (Experiments 1-3) were conducted online to facilitate large sample sizes (to increase the reliability and reproducibility of the results). Extensive research conducted in our laboratory has consistently demonstrated that online data can match the quality of data collected in traditional laboratory settings[30,97-102] provided measures are taken to ensure standardized sound presentation, encourage participant compliance, and promote active engagement in the tasks. Online participants were recruited using the Prolific platform. To ensure data quality, participants were prescreened to have at least a 95% submission approval rate, and to have not completed any online experiment hosted by the authors' lab in the past 6 months (to avoid familiarity with our experimental stimuli and task). All online participants were instructed to complete the experiment in a quiet environment and completed a headphone check experiment[103] prior to the main experiment to help maximize sound presentation quality. Participants adjusted the volume of a calibration sound to a comfortable level at the start of the experimental session, and all stimuli were scaled relative to this maximum level. Participants also completed 12 catch trials in each experiment. These were intended to make sure participants were paying attention to the experiment and served as an additional screening metric. Catch trials presented isolated words (spoken by one of the authors) in silence using the same clip for cue and mixture and were randomly intermixed with experimental trials. Participants scoring less than 11 correct catch trials (91.6% accuracy) were excluded from our analysis.

Experiments 6 and 7 measuring the effects of spatial separation on attentional selection were run in-person over a speaker array (detailed below) to ensure the accuracy of stimuli playback location. Because participants were monitored for the duration of these experiment, catch trials were not included.

Informed consent

All participants provided informed consent and the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (COUHES) approved all experiments.

Experiment 1: effect of distractors on attentional selection in monaural conditions

Experiment 1 measured cued word recognition as a function of signal-to-noise ratio (SNR) and distractor type. Talkers and speech materials were not included in model training so as to enable

fair comparisons between models and humans. Nine distractor types were used: 1-talker same-sex, 1-talker different-sex, 2-talker, 4-talker, 8-talker babble, 1-talker Mandarin speech, stationary noise, recorded auditory scenes, and instrumental music.

Stimuli

Cue, target, and English-speech distractor signals were sourced from the portion of Spoken Wikipedia Corpora[104] material screened in Feather et al[95]. Feather et al. removed Spoken Wikipedia articles due to a) potentially offensive content for human listening experiments, b) missing data or c) having bad audio quality (for example, due to computer generated voices of speakers reading the article or the talker changing mid-way through the clip). We applied further screening to this set to identify clips centered on words in the model vocabulary (316,748 clips), so that the model could also be run on the experiments.

Target clips were additionally constrained to be sex-balanced. The in-vocabulary clips were further screened to find words with at least one example produced by both a male and female talker, yielding 488 target words. One male- and one female-talker example was uniformly sampled for each target word, giving 976 total target clips. Cue clips were drawn from target excerpts centered on words not included in the model vocabulary.

English speech clips for the 1-, 2-, and 4-talker distractor conditions were sampled from the full set of Spoken Wikipedia excerpts containing words in the model vocabulary. This made it possible to measure the rate of distractor word reports without needing to reuse target clips as distractors in a given experiment. So that the effect of talker-sex similarity could be analyzed in the 1-talker distractor condition, one male and one female distractor clip were sampled for each target clip. Two 2-talker distractor clips were sampled for each target clip by summing either two male or two female clips. The 4-talker distractor clips were created by summing the two male and two female distractor clips. The talker identity and middle word of the distractor clip(s) were constrained to differ from those of the target clip. The 1-talker Mandarin distractor clips were sourced from the Mandarin validation (termed “dev” in the dataset documentation) portion of Common Voice (version 9). One male and one female Mandarin speech clip were sampled for each target clip to be included in the talker sex analysis. Speech-shaped noise was synthesized for each target clip by imposing the power spectrum of the target clip on white noise. Instrumental music, auditory scenes, and 8-talker babble were sampled from the MUSDB18, IEEE AASP CASA Challenge, and Common Voice test clips used in Saddler & McDermott (2024) (450, 400, 400 total clips respectively).

Target clips were combined with distractor clips from each condition at 6 SNRs (infinite, i.e. no-distractor, and -9, -6, -3, 0, +3 dB) producing 44,896 possible mixture stimuli (976 target clips with no distractor + 976 target clips x 9 distractor types x 5 SNRs). Cue, target, and distractor signals were 2-seconds in duration, sampled at 44,1 kHz. Mixtures were obtained by setting the target clip against the distractor clip at the given SNR. All cue and mixture clips were normalized to a RMS of 0.02 for model experiments, and were presented at the same participant-determined level for the human experiments (described in the Procedure section).

Procedure

Individual participants completed 12 catch trials and 184 experimental trials (4 trials x 5 SNRs x 9 distractor conditions + 4 no-distractor trials). Each participant heard a random sample of 184 of the 488 target words, with the talker sex randomly determined for each word. The words were then randomly assigned to distractor and SNR conditions for each participant, constrained to yield both a sex balance over target talkers and target-distractor sex conditions, and unique cue and

mixture clips. Target-distractor pairings were constrained as described above. The 12 catch trials were randomly intermixed with the 184 experimental trials.

On every trial, participants first heard the 2-second cue, then a 0.5-second delay, then the 2-second mixture. This audio sequence was initiated by a button click. Participants reported the word said by the target voice in the middle of the mixture clip (defined as the word overlapping the 1-second mark of the mixture). To equate the task for humans and models, participants were asked to select words from a list of the 800 words in the model vocabulary. Participants typed responses into a text box and, while they typed, the web page displayed a list of 800 words that was continuously updated to only include matches to the string they typed. Only words in the word list could be entered, analogous to how the models could only report one of these 800 words. Participants received feedback after each trial, displaying the trial outcome ("correct" or "incorrect"), the number of remaining trials, and their running accuracy. On incorrect trials, the correct word was also displayed to help participants lean to perform the task.

The experiment was run online over the Prolific platform. Participants first heard a calibration sound and adjusted the presentation level of their computer to be comfortable prior to the start of the experiment, and then completed a headphone check[103] to help ensure good sound quality. Only participants who passed the headphone check completed the main experiment. We then only analyzed data from participants who responded correctly to at least 91% of catch trials and who self-reported normal hearing. 195 participants (98 female, 92 male, 4 non-binary, 1 no-report) met these criteria. Participants ages were between 18 and 71 (median 33) years.

Model experiment

Models were tested on all combinations of the 976 target clips (both male and female examples of each target word) with the 5 SNRs, 9 distractor conditions, and no-distractor condition (44,896 total stimuli) used in the human experiments. Because listeners used headphones rather than free-field speakers, audio signals given as input to the model were not run through the room simulator. Instead, each mono clip was presented diotically (i.e., the mono stimulus waveform was input for both the left and right channel of the model).

Analysis

To account for the possibility that participants correctly recognized the target speech signal but inadvertently reported a word that was adjacent to the trial's target word rather than the target word itself, word recognition performance was measured by contrasting a participant's response against all in-vocabulary words contained in a trial clip. In all conditions, a response was counted as correct if it matched any of the in-vocabulary words that were present in the target utterance. In the 1-distractor condition, confusions were analogously measured as reports of words contained in the transcript of the distractor utterance. This scoring was applied to both participant and model reports.

To analyze the effects of target-distractor sex similarity and language familiarity, we reanalyzed both the English and Mandarin 1-distractor conditions for both participants and models. To analyze target-distractor sex similarity, trials from both distractor language conditions were first pooled, then the mean for each distractor sex condition was obtained for each participant (or model architecture). To analyze the effect of distractor language, trials from both distractor sex conditions were pooled within language, then the mean for each distractor language condition was obtained for each participant (or model architecture).

Experiment 2: effect of harmonicity on attentional selection in monaural conditions

Experiment 2 measured how violations of harmonicity (the tendency of frequency components to be integer multiples of a fundamental frequency) impact cued word recognition. Participants were presented with a cue, followed by either the target excerpt alone, or the target excerpt mixed with a distractor excerpt. Harmonic, inharmonic, and whispered stimulus variants were generated as in Popham et al., 2018 (summarized below), using the stimuli from the English 1-distractor condition of Experiment 1. Twelve target-distractor harmonicity conditions were used (3 target harmonicities – harmonic, inharmonic, and whispered – x 3 distractor harmonicities + 3 target harmonicities with no distractor). We only analyzed the conditions in which the target and distractor were of the same type (harmonic/harmonic, inharmonic/inharmonic, and whispered/whispered), to be consistent with the original experiment run by Popham et al. Participant and model responses were analyzed identically to Experiment 1.

Stimuli

Experiment 2 used the set of cue, target, and distractor pairings from the 1-talker same-sex and 1-talker different-sex conditions of Experiment 1. All stimuli were resynthesized from the original speech excerpts of Experiment 1 using the STRAIGHT algorithm[105,106]. Inharmonic speech examples were synthesized by shifting harmonic frequency components above the fundamental frequency by either -30% or 30% (which maximally reduced intelligibility in Popham et al., 2018). Jitter values were sampled independently for each harmonic's frequency of each original speech clip but were constrained (via rejection sampling) such that adjacent harmonics were always separated by at least 30 Hz. To generate the stimuli for an inharmonic trial, one jitter pattern was sampled for the target excerpt, and this same jitter pattern was applied to the cue and distractor utterances. Harmonic stimuli were synthesized by running stimuli through STRAIGHT without changing any of the constituent harmonic frequencies. Whispered speech was created as in Popham et al. 2018, and the description of the synthesis from that paper is reproduced here with minor edits for clarity. Whispered stimuli were generated by omitting the sinusoidal excitation used for harmonic/inharmonic synthesis and high-pass filtering the noise excitation to simulate breath noise in whispered speech. The filter was a second-order high-pass Butterworth filter with a (3 dB) cutoff at 1200 Hz whose zeros were moved toward the origin (in the z-plane) by 5%. The resulting filter produced noise that was 3 dB down at 1600 Hz, 10 dB down at 1000 Hz, and 40 dB down at 100 Hz, which to the authors sounded like a good approximation to whispering. Without the zero adjustment, the filter removed too much energy at the very bottom of the spectrum. The noise excitation was combined with the time-varying spectral envelope using the same procedure employed for harmonic and inharmonic speech. The noise-excited stimuli were thus generated from the same spectrotemporal envelope used for harmonic and inharmonic speech, just with a different excitation signal.

Experiment 2 used the same combinations of cue, target, and distractor source clips as Experient A. Harmonic, inharmonic, and whispered versions of each target-distractor pairing were synthesized as described above. To produce the 9 target-distractor harmonicity combinations, each version of a target clip was crossed with each version of the paired distractor clip. Cue clips always used the harmonicity of the target clip. Signal sampling rates and presentation levels matched those of Experiment 1.

Procedure

The procedure of Experiment 2 was identical to that of Experiment 1, with the following exceptions. Participants each completed 12 catch trials and 180 experimental trials (15 trials x 12 harmonicity conditions). Each participant heard a random sample of 180 of the 488 target words, with the harmonicity condition for each word randomly assigned for each participant. The 12 catch trials were randomly intermixed with the 180 experimental trials. 90 participants (43 female, 47

male) met our inclusion criteria (scoring 91% on the catch trials). Participant ages were between 19 and 64 (median 34.5) years. We note that this experiment used a different type of cueing than was used in the original experiment of Popham et al. (the Popham et al. experiment showed participants the first two words of the sentence they were supposed to listen to, prior to the start of the audio stimulus). This task difference may be responsible for the lower rate of confusions we observed compared to the Popham et al. experiment.

Model experiment

Models were tested on all combinations of the 976 target clips (both male and female examples of each of the 488 target words) with the 9 harmonicity conditions, and 3 no-distractor conditions (23,424 total stimuli) used in the human experiments. As in Experiment 1, the mono audio clips were not run through a room simulator, and were instead presented to the model diotically.

Experiment 3: word recognition in naturalistic auditory textures

To obtain a stronger test of the extent to which the models captured human speech-in-noise perception, we ran our models on an experiment from Saddler and McDermott 2024[22], which probed human speech-in-noise recognition across a large set of naturalistic distractor signals. The description of the human experiment is reproduced from the original paper with minor edits for brevity.

Human word recognition accuracy was measured in 43 different naturalistic auditory textures. 376 speech excerpts from the evaluation portion of the Word-Speaker-Noise dataset[95] were embedded in 376 unique exemplars of each auditory texture. The 2s texture exemplars were previously generated[37] to match the statistics of 43 recorded real-world textures[107]. The success of the iterative synthesis algorithm[107] used to generate the textures 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). Speech excerpts were randomly assigned to one of the 43 texture conditions with the SNR fixed at −3 dB (each participant heard a different random assignment). The experiment was run online and included 47 participants (24 female, 23 male) who self-reported normal hearing, passed a headphone check[103], 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 old.

Model experiment

We simulated a version of the original human experiment, modified to make the task compatible with the model task. For each target excerpt, a corresponding cue was obtained by sourcing a different excerpt of the target talker from the Word-Speaker-Noise dataset. We then measured model word recognition accuracy for speech embedded in each of 43 auditory textures, using the same stimuli as in the human experiment. Models were evaluated on the full stimulus set (16,168 stimuli = 376 speech excerpts × 43 auditory textures).

Experiment 4: spatial tuning for masked speech

We simulated experiment 1 of Byrne et al. (2023)[34], which measured spatial tuning for the release of masking of target speech, manipulating the distractor type and proximity to the target source. The original experiment used a closed-set masked speech identification task[108], in which listeners were presented with 5-word sentences, in English with fixed syntactic structure, spoken by one of 12 female talkers (stimuli fully described in Kidd et al.[108]). Listeners were cued to the target stream by being told the starting word of the target sentence, which was fixed throughout the experiment, and that the target would occur directly in front of their position. Three distractor conditions were used: 1) two competing female talkers drawn from the 11 remaining talkers, 2)

time-reversed speech made from the same sources used in condition 1, and 3) speech-shaped speech-envelope-modulated noise derived from the speech distractors of condition 1. The target source was randomly selected per trial, and was always located at $0°$ elevation and $0°$ azimuth. Two distractor sources were positioned symmetrically in azimuth around the target talker at angles of $\pm 0°, \pm 5°, \pm 10°, \pm 20°$, and $\pm 40°$ azimuth and $0°$ elevation. For each distractor x angle condition, speech reception thresholds were measured using a one-down-one-up adaptive procedure, varying the signal-to-noise ratio between the target and summed distractor signals to estimate the 50% correct point on the psychometric function. The distractor level was held constant such that the sum of the two distractors was 60 dB, and the target level was adjusted by 3dB increments depending on participant accuracy. Trials were evaluated as correct, and the target level was reduced, if participants correctly identified the remaining 4 words in the target sentence. We reproduced the first of the three conditions (the one with two competing talkers) in our model experiment. Data was collected for 18 normal hearing native English speaking listeners (11 female, 6 male, 1 non-specified; aged 18-40 years). We scanned the results in from the original paper figure.

Model experiment

We simulated a version of the original human experiment using a virtual anechoic room and the same-sex distractor speech materials from Experiment 7. The model simulation differed in two respects from the human experiment. First, the target voice was designated by cue that was a different excerpt of the target talker (as in the other model experiments in this paper), rather than by one of the words spoken by the target talker. Second, in lieu of the adaptive procedure, we ran the model on 9 signal-to-noise ratios (-18 dB to 6 dB SNR in 3 dB steps) at each spatial location. All combinations of 976 stimuli (cue, target, and 1-talker same-sex-distractor pairs), 5 loudspeaker presentation conditions ($\pm 0°, \pm 5°, \pm 10°, \pm 20°$, and $\pm 40°$ azimuth), and 9 SNRs were simulated (43,920 total stimuli). Model thresholds for each loudspeaker condition were estimated as follows. First, word recognition performance was measured at each SNR and presentation condition. Second, thresholds at each presentation condition were estimated by fitting a second-order polynomial to the performance by SNR curve, and calculating the SNR granting 50% of the model's maximum performance.

Experiment 5: the precedence effect with concurrent speech signals

We simulated experiment 1 of Freyman et al. (1999)[38], which measured the benefit of perceived spatial separation on speech intelligibility, using the precedence effect to induce illusory spatial separation between concurrent talkers. The "precedence effect" traditionally refers to a perceptual phenomenon that occurs when two sounds played in quick succession from different locations seem to originate from the first location[35,36]. In this experiment, a target talker was superimposed with a distractor talker. The experiment included a condition where the target and distractor were co-located and a condition where they were spatially separated, as well as conditions where the locations were illusorily shifted by adding a second copy of one of the signals at a different location, with a small temporal offset.

In the original experiment, 56 participants (sexes not reported) listened to recordings of speech, with one target talker and one distractor, and reported the keywords spoken by the target talker. Speech materials were of syntactically correct sentences that had no meaning. The target talker was always the same female voice, while the distractor was always the same alternate female voice. The experiment took place in an anechoic chamber, with loudspeakers positioned in a semicircular arc at 0° and 60° to the right relative to the listener position, at a distance of 1.9 meters from the listener. Six conditions were used for target and distractor locations (numbers reflect ordering in the original paper): 1) both target and distractor presented at 0°, 2) target presented at 0° with distractor presented at 60°, 3) target presented at 0° with distractor illusorily

presented at 0° via precedence effect, 4) target presented at 0° with distractor illusorily presented at 60° via precedence effect, 5) both target and distractor illusorily presented at 0° via precedence effect, and 6) target illusorily presented at 0° and distractor illusorily presented at 60°. Illusory separation in conditions 3-6 was enabled by presenting signals from both positions and starting playback at the "perceived" location 4-ms before starting playback at the "lag" location (taking advantage of the precedence effect). Four signal-to-noise ratios (SNR) were presented per loudspeaker condition: -12 dB, -8 dB, -4 dB, and 0 dB SNR. We reproduce conditions 1, 2, and 4.

Model experiment

We simulated the original human experiment using a virtual anechoic room and the speech materials from Experiment 7. All combinations of 1,952 stimuli (cue, target, and single-distractor pairs), 3 loudspeaker presentation conditions (conditions 1, 2, and 4 described above), and the four original SNRs (-12 dB, -8 dB, -4 dB, and 0 dB SNR) were simulated (23,436 total stimuli). To simulate the precedence effect, we first spatialized the distractor to the lead channel. Next, we applied the 4-ms onset delay (by zero-padding) to the distractor, then spatialized it to the "lag" location. Finally, the two binaural waveforms were summed in each channel.

Experiment 6: effect of spatial separation in azimuth and elevation

The benefit of target-distractor spatial separation was measured for separation in azimuth and elevation. Stimuli in Experiment 6 were a subset of the speech clips used in Experiments 1 and 3 that were screened to be suitable for sound localization experiments (described below). Experiment 6 was conducted in person, and stimuli were presented over a loudspeaker array (described below). Participant and model responses were analyzed as in Experiment 1

Stimuli

Speech excerpts for Experiments 6 and 7 were sourced from the set of Spoken Wikipedia Corpus examples used in the online monaural experiments (Experiments 1 and 3). Cue, target, and distractor stimuli were screened to include excerpts with frequency content beyond 16kHz to help enable sound localization in elevation. Target excerpts were further screened under the constraint that each word was spoken by both a male and female talker. These screening steps reduced the number of unique target words to 488 from 800, (976 total target excerpts, 488 per talker sex). This vocabulary size was large enough to allow for one unique word per trial in the in-person experiment. Cue excerpts were selected for each target excerpt by finding clips of the target talker centered on words not in the target vocabulary. To allow for analysis of talker-distractor sex similarity effects, each target excerpt was paired with both a same-sex and different-sex distractor. To enable use of symmetrically positioned distractors in the azimuth conditions, two distractors per sex condition were sampled from the screened set of distractors. This screening resulted in 1,952 unique combinations of cue, target, and two-talker distractors that were sex-balanced across both target talkers and target-distractor sex pairings.

Speaker array

All experiments measuring effects of spatial separation were run using a 19-by-5 array of loudspeakers arranged on a hemisphere (2 meter radius). Participants were seated at the center. Relative to the participant's head, the array spanned 180° in azimuth (frontal hemifield) and -20° to 40° in elevation, with speakers spaced every 10° in both azimuth and elevation. Speaker locations were coded using labels that were attached to the bottom of each loudspeaker. Labels were formatted as a combination of one letter followed by two digits (e.g. A10), with letters indicating elevation (A-G for 40° to -20° degrees elevation) and numbers indicating azimuth (1-19 indicating 90° to -90° elevation). Participant responses were collected on an Apple iPad held by the participant during the experiment. Participants were instructed to direct their head at the

loudspeaker directly in front of them for the duration of the stimulus. Once the stimulus ended, participants could look at the iPad to enter their response. Participants were instructed to redirect their head towards the front loudspeaker before triggering the start of the next trial. Compliance with these instructions was confirmed by experimenter observation.

Procedure

The trial procedure and task were identical to that of Experiment 1 (which was run online), except that participants heard sounds from loudspeakers rather than headphones/earphones, and that the cue/target/distractor varied in spatial position. First, a 2-second cue excerpt would play from a specific location, allowing participants to hear both the target talker's voice and location. After a 0.5-second delay, a 2-second target excerpt was played from the same location, while two 2-second distractor excerpts were played simultaneously at locations that varied with respect to the target location. The target and distractor were separated in either azimuth or elevation, depending on the condition.

Two distractors were used in every trial. When the distractors varied in azimuth, the two (symmetrically placed) distractors served to preserve the signal-to-noise ratio at each ear across different target-distractor separations[34,109]. When the distractors varied in elevation, the use of two distractors served to match the distractor properties to those of the azimuth conditions, enabling a more controlled comparison. Listeners reported the middle word spoken by the target talker (i.e. the word overlapping the 1-second mark). Participants entered responses by typing on an iPad, connected to a local server running the same experimental interface used in the online experiments.

A common set of target positions and distractor offsets were used in azimuth and elevation conditions to compare the effect of spatial separation in each dimension. Distractors were offset relative to target positions by either 0 (i.e., co-located), 10, or 60 degrees in either azimuth or elevation. These offsets were chosen as those likely to reveal the extent of spatial tuning subject to the limits of our speaker array: 10 degrees was the smallest offset between speakers, while 60 degrees was the largest span available in elevation (minimum elevation of -20 degrees; maximum of 40 degrees). Cue and target excerpts were always presented at 0 azimuth, and at either -20- or 40-degrees elevation (so that 60-degree target-distractor offsets could be used). To equate the uncertainty of the target location in azimuth and elevation, one of the two possible target elevations was sampled randomly for each participant, and was used for all trials (so that for any single participant the target was always at the same location in azimuth and elevation).

In the conditions in which target and distractors were separated in azimuth, the cue, target, and distractor signals shared a common elevation. The two distractor excerpts were presented symmetrically around the target at the specified offset. In the conditions in which target and distractors were separated in elevation, cue and target excerpts were always presented at 0 degrees azimuth. The two distractor excerpts were presented from the same position, at 0 azimuth and an elevation offset from the target elevation by the specified amount. When the target was at -20 degrees the distractors were offset above the target location; when the target was at 40 degrees the distractors were offset below the target location.

The experiment used the method of constant stimuli. At each target-distractor position, word recognition was measured at a set of signal-to-noise ratios (-9, -6, -3, 0, 3, and 6 dB). In trials with co-located distractors, the two distractor waveforms were first summed, and then their combined signal was normalized to 65 dB SPL. In trials with symmetrically positioned distractors, each distractor was individually set to 61.99 dB SPL, such that the level of the two distractors was 65 dB SPL on average when summed at the ear. The level of the target excerpt was then adjusted

to obtain the desired SNR as 65 dB + SNR. The cue signal was always presented at 65 dB SPL, so that cue level would not be informative for the task.

Participants completed 16 trials for each combination of target position, distractor offset, and SNR, totaling 480 trials (1 target position x 5 distractor offsets x 6 SNRs x 16 trials). Trials were randomly ordered and then grouped into six blocks of 80 trials, with participants encouraged to take self-timed breaks after each block. The entire experiment took approximately 2-hours to complete. The selection of stimuli and assignment of excerpt to condition were randomized independently for each participant. 33 participants (22 female, 10 male) completed the experiment. Participant ages were between 18 and 40 years (median age = 23).

Analysis

Speech reception thresholds were estimated for each distractor offset in azimuth or elevation using the following procedure. First, word recognition performance as a function of signal-to-noise ratio was measured for each participant as in Experiment 1, at each distractor offset. Second, thresholds at each distractor offset and offset direction were estimated by fitting a second-order polynomial to the mean of these values across participants and calculating the SNR granting 50% performance from this curve. The uncertainty in these thresholds was estimated via bootstrap (thresholds were measured from 10,000 bootstrap samples of 33 participants; results graphs plot the mean and standard deviation of these thresholds).

Model experiment

Models were tested at all combinations of the 1,952 stimuli (all cue, target, and two-distractor pairings) with the 5 distractor offsets and 6 SNRs (58,560 total combinations). Stimuli were spatialized using a virtual rendering of the loudspeaker array room participants were tested in. Sound pressure levels in the model experiment were identical to the human experiment. To approximately equate the model and human testing conditions, 30dB of pink noise was added to the model stimuli to match the background noise level present in the loudspeaker array room used to test participants. Model speech reception thresholds for each distractor offset and direction were estimated by bootstrapping over the 10 model architectures (10,000 samples).

Experiment 7: measuring the width of spatial attention

Experiment 7 measured how spatial release from masking in azimuth depends on the azimuthal position of the attended (target) talker relative to the listener. Eight spatial configurations were used: 2 target azimuths (0° or 90°) x 4 distractor azimuthal offsets (0°, 10°, 30°, or 90°). Experiment 7 used the same cue, target, and distractor clips as Experiment 6, except that only one distractor talker was included on each trial. All stimuli were presented at 65 dB SPL, with the target and distractor signal-to-noise ratio set to 0 dB. Participant and model responses were analyzed as in Experiment 1.

Procedure

Experiment 7 used the same procedure as Experiment 6, except for the following changes. Participants completed 160 total trials (2 target positions x 4 distractor offsets x 20 trials). Each trial had a different target word. The 160 target words for a participant were randomly sampled from the full set of 488 target words, and their assignment to spatial locations was randomized per participant. Trials were arranged into 4 blocks of 40 trials. All the trials in a block had the same target position, and the block included 10 trials of each distractor offset. The block order and the order of trials within blocks was randomized per participant. 28 participants (16 female, 12 male) completed the experiment. Participant ages were between 19 and 39 years (median age = 26).

Because the speaker array did not enable left-right symmetric offsets at the 90° azimuth positions (the array did not extend beyond 90° on either side), we used the following procedure to better equate the uncertainty over source positions between the 0° and 90° target conditions. First, distractors were offset in a consistent direction relative to the target (i.e., distractors occurred either to the left or right of the target for the entire experiment). The distractor offset direction was alternated across participants. Second, the 0° and 90° target conditions were achieved by rotating participants to face either the center or the side of the array, while maintaining the source playback locations. In all trials, the cue and target were presented at the center of the array, with distractors occurring 0°, 10°, 30°, or 90° away from the target in azimuth (either to the participant's left or right depending on the offset direction). In blocks testing the 0° target condition, participants faced the center speaker. In blocks testing the 90° target condition, participants were rotated so the center speaker was at 90° relative to the participant, and the distractor offsets were positioned to the same side as in the 0° target condition. All stimuli were presented at 0° elevation.

Model experiment

Models were tested at all combinations of the 1,952 stimuli (cue, target, and single-distractor pairs) with the 2 target positions and 4 distractor offsets (15,616 total combinations). All model stimuli were RMS normalized to 0.02. The room simulation and inclusion of noise were identical to Experiment 6.

Analysis of model locus of attention (Figure 5)

To analyze the locus of attention in our models, we compared the model's representation of a single-talker target clip with its representation of a mixture composed of the target and a single-talker distractor clip, as a function of whether the model was cued to the target talker or distractor talker. The logic for this analysis was that the representations of a single source and a mixture containing that source would be similar if attentional selection for the single source's talker was applied to the mixture. We analyzed the similarity between targets and mixtures containing one distractor talker, presented signals diotically to the model, and quantified similarity using Pearson's correlation coefficient. The cue, target, and single-distractor speech excerpts curated for Experiment 1 were used as stimuli. The following procedure was performed for each cue, target, and one-distractor pair. First, representations of the isolated target clip, $x_{target}^{(l)}$, distractor clip, $x_{distractor}^{(l)}$, and their mixture, $x_{mixture}^{(l)}$, were obtained separately from each layer, l, using the same cue to designate the target talker for each. Second, we measured Pearson's correlation coefficient between target and mixture as $\mathrm{corr}\left(x_{target}^{(l)}, x_{mixture}^{(l)}\right)$, and between distractor and mixture as $\mathrm{corr}\left(x_{\mathrm{distractor}}^{(l)}, x_{mixture}^{(l)}\right)$ where $\mathrm{corr}(x, y)$ is the Pearson's correlation:

$$\mathrm{corr}(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

and i indexes the flattened feature dimension of size $N =$ Channels × Frequency × Time. 1000 cue, target, and single-distractor pairs were used to measure correlations (500 same-sex and 500 different-sex distractors). Targets and distractors were presented at 0 dB SNR, and all signals were RMS-normalized to 0.02.

Aggregate measures of human-model similarity (Figure 6b&c)

Human-model behavioral similarity was quantified by comparing human and model performance in each experimental condition, separately for each model. To ensure that our conclusions were robust to the choice of similarity metric, the following analysis was performed using both Pearson's correlation coefficient and root-mean-squared-error (RMSE) as similarity metrics. For

each model, we measured the similarity between the mean human behavior (averaged across experiment participants) and mean model behavior, using means computed per experimental condition (e.g. a particular distractor type x signal-to-noise ratio). For the feature-gain model, the mean model behavior was taken as the average across the 10 network architectures. The statistical significance of the effect each alternative architectural constraint (e.g. early-only, late-only, or attention-free architectures) had on overall human-model similarity was assessed by comparing the human-model similarity scores against the distribution of the feature-gain model's human-model similarity scores via sign test. Two-tailed p values are reported, and effect sizes were quantified by measuring the difference between similarity scores over the mean score of the feature-gain model.

Statistics

Statistical tests: Interactions were assessed using a repeated measures Analysis of variance (ANOVA), implemented using the statsmodels[110] and Pingouin[111] packages for python. Significance of the effect of model architectural constrains on human-model similarity used sign tests implemented via statsmodels and scipy.stats packages. Paired two-tailed t-tests used to measure the effect of distractor harmonicity on model performance were implemented using the scipy.stats[112] packages.

Experiment 6 did not permit a conventional test for an interaction because thresholds were difficult to estimate reliably in individual participants. Instead, we assessed the statistical significance of the interaction between the direction of offset (azimuth or elevation) and offset magnitude (0°, 10°, or 60°) using a permutation test. First, the interaction between conditions was measured by subtracting off the marginal contribution of each independent variable:

$$\text{interaction}_{ij} = \text{threshold}_{ij} - \mu^{(i)}_{\text{direction}} - \mu^{(j)}_{\text{offset}} + \mu_{\text{total}}$$

where i indexes the direction of offset (azimuth or elevation), j indexes the offset magnitude, $\mu^{(i)}_{\text{direction}}$ is the average threshold over offsets for direction i, $\mu^{(j)}_{\text{offset}}$ is the average threshold over directions for offset j, and μ_{total} is the average over both direction and offset. An overall interaction effect was then obtained as the sum of squares of these interaction terms:

$$\text{overall interaction} = \sum_{i,j} \text{interaction}^2_{ij}$$

The overall interaction was re-computed 10,000 times with direction of offset and offset magnitude labels permuted at the subject level, obtaining a null distribution used to calculate a p-value for the actual overall interaction.

Error bars: Except where otherwise noted in figure captions, error bars in results figures indicate ± 1 standard error of the mean (SEM) across experiment participant results (human results) or across the 10 network architectures (model results).

Acknowledgments

Research supported by National Institutes of Health grant R01 DC017970. The authors thank Adanna Thomas for running the human participants in Experiments 6 and 7.

References

1 Cherry, E. C. Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America* **25**, 975-979 (1953).

2 Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18**, 193-222 (1995).

3 Driver, J. A selective review of selective attention research from the past century. *British Journal of Psychology* **92**, 53-78 (2001).

4 Maunsell, J. H. R. & Treue, S. Feature-based attention in visual cortex. *Trends Neurosci.* **29**, 317-322 (2006).

5 Fritz, J. B., Elhilali, M., David, S. V. & Shamma, S. A. Auditory attention — focusing the searchlight on sound. *Curr. Opin. Neurobiol.* **17**, 437-455 (2007).

6 Shinn-Cunningham, B. G. Object-based auditory and visual attention. *Trends Cogn. Sci.* **12**, 182-186 (2008).

7 Chun, M. M., Golomb, J. D. & Turk-Browne, N. B. A taxonomy of external and internal attention. *Annual Review of Psychology* **62**, 73-101 (2011).

8 Carrasco, M. Visual attention: The past 25 years. *Vision Research* **51**, 1484-1525 (2011).

9 Motter, B. C. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience* **14**, 2178-2189 (1994).

10 Treue, S. & Martinez Trujillo, J. C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579 (1999).

11 McAdams, C. & Maunsell, J. H. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience* **19**, 431-441 (1999).

12 Martinez Trujillo, J. C. & Treue, S. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology* **14**, 744-751 (2004).

13 Martinez Trujillo, J. C. & Treue, S. Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron* **35**, 365-370 (2002).

14 Reynolds, J. H. & Heeger, D. J. The normalization model of attention. *Neuron* **61**, 168-185 (2009).

15 Kerlin, J. R., Shahin, A. J. & Miller, L. M. Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *Journal of Neuroscience* **30**, 620-628 (2010).

16 Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233-236 (2012).

17 Ding, N. & Simon, J. Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences* **109**, 11854-11859, doi:10.1073/pnas.1205381109 (2012).

18 Zion-Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D. & Schroeder, C. E. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **77**, 980-991 (2013).

19 Puvada, K. C. & Simon, J. Z. Cortical representations of speech in a multitalker auditory scene. *Journal of Neuroscience* **37**, 9189-9196 (2017).

20 Puschmann, S., Regev, M., Fakhar, K., Zatorre, R. J. & Thiel, C. M. Attention-driven modulation of auditory cortex activity during selective listening in a multispeaker setting. *Journal of Neuroscience* **44**, e1157232023 (2024).

21 Rajalingham, R., Issa, E., Bashivan, P., Kar, K., Schmidt, K. & DiCarlo, J. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* **38**, 7255-7269 (2018).

22 Saddler, M. R. & McDermott, J. H. Models optimized for real-world tasks reveal the necessity of precise temporal coding in hearing. *Nature Communications* **15**, 10590, doi:10.1038/s41467-024-54700-5 (2024).
23 Bronkhorst, A. W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica* **86**, 117-128 (2000).
24 McDermott, J. H. The cocktail party problem. *Current Biology* **19**, R1024-R1027, doi:10.1016/j.cub.2009.09.005 (2009).
25 Broadbent, D. E. Listening to one of two synchronous messages. *The Journal of Experimental Psychology* **44**, 51-55 (1952).
26 Brungart, D. S. Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* **109**, 1101-1109 (2001).
27 Kidd, G., Arbogast, T. L., Mason, C. R. & Gallun, F. J. The advantage of knowing where to listen. *Journal of the Acoustical Society of America* **118**, 3804-3815, doi:10.1121/1.2109187 (2005).
28 Best, V., Ozmeral, E. J. & Shinn-Cunningham, B. G. Visually-guided attention enhances target identification in a complex auditory scene. *The Journal of the Association for Research in Otolaryngology* **8**, 294-304 (2007).
29 Broadbent, D. E. Failures of attention in selective listening. *The Journal of Experimental Psychology* **44**, 428-433 (1952).
30 Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. & 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, doi:10.1016/j.neuron.2018.03.044 (2018).
31 Van Engen, K. J. & Bradlow, A. R. Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America* **121**, 519-526 (2007).
32 Calandruccio, L., Dhar, S. & Bradlow, A. R. Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America* **128**, 860-869 (2010).
33 Popham, S., Boebinger, D., Ellis, D. P., Kawahara, H. & McDermott, J. H. Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications* **9**, 2122 (2018).
34 Byrne, A. J., Conroy, C. & Kidd, G. Individual differences in speech-on-speech masking are correlated with cognitive and visual task performance. *Journal of the Acoustical Society of America* **154**, 2137-2153 (2023).
35 Litovsky, R. Y., Colburn, H. S., Yost, W. A. & Guzman, S. J. The precedence effect. *Journal of the Acoustical Society of America* **106**, 1633-1654, doi:10.1121/1.427914 (1999).
36 Brown, A. D., Stecker, G. C. & Tollin, D. J. The precedence effect in sound localization. *Journal of the Association for Research in Otolaryngology* **16**, 1-28, doi:10.1007/s10162-014-0496-2 (2015).
37 Francl, A. & McDermott, J. H. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behavior* **6**, 111-133 (2022).
38 Freyman, R. L., Helfer, K. S., McCall, D. & Clifton, R. K. The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America* **106**, 3578-3587 (1999).
39 Worley, J. W. & Darwin, C. J. in *The International Conference on Auditory Display* (Kyoto, Japan, 2002).
40 McAnally, K. I., Bolia, R. S., Martin, R. L., Eberle, G. & Brungart, D. S. in *The Human Factors and Ergonomics Society 46th Annual Meeting* 588-591 (2002).

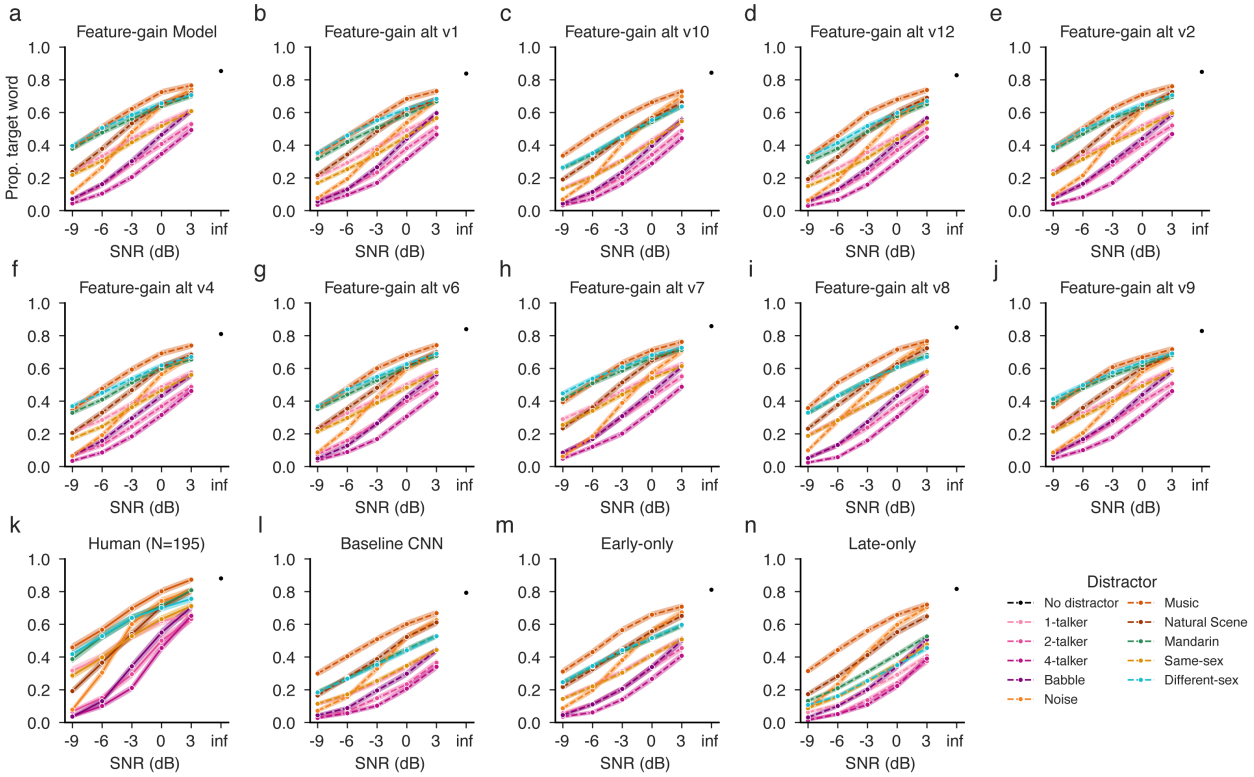
41 Best, V. *Spatial hearing with simultaneous sound sources: A psychophysical investigation* PhD thesis, University of Sydney, (2004).
42 Martin, R. L., McAnally, K. I., Bolia, R. S., Eberle, G. & Brungart, D. S. Spatial release from speech-on-speech masking in the median sagittal plane. *Journal of the Acoustical Society of America* **131**, 378-385 (2012).
43 Berwick, N. & Lee, H. Spatial unmasking effect on speech reception threshold in the median plane. *Applied Science* **10**, 5257 (2020).
44 González-Toledo, D., Cuevas-Rodríguez, M., Vicente, T., Picinali, L., Molina-Tanco, L. & Reyes-Lecuona, A. Spatial release from masking in the median plane with non-native speakers using individual and mannequin head related transfer functions. *Journal of the Acoustical Society of America* **155**, 284-293 (2024).
45 Sandel, T. T., Teas, D. C., Feddersen, W. E. & Jeffress, L. A. Localization of sound from single and paired sources. *Journal of the Acoustical Society of America* **27**, 842-852, doi:10.1121/1.1908052 (1955).
46 Mills, A. W. On the minimum audible angle. *Journal of the Acoustical Society of America* **30**, 237-246, doi:10.1121/1.1909553 (1958).
47 Wood, K. C. & Bizley, J. K. Relative sound localisation abilities in human listeners. *Journal of the Acoustical Society of America* **138**, 674-686, doi:10.1121/1.4923452 (2015).
48 Shinn-Cunningham, B. G. in *International Conference on Auditory Display.* 126-134.
49 Brandewie, E. & Zahorik, P. Prior listening in rooms improves speech intelligibility. *Journal of the Acoustical Society of America* **128**, 291-299 (2010).
50 Petkov, C. I., Kang, X., Alho, K., Bertrand, O., Yund, E. W. & Woods, D. L. Attentional modulation of human auditory cortex. *Nature Neuroscience* **7**, 658-663 (2004).
51 Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology* **28**, 3976-3983 (2018).
52 O'Sullivan, J., Herrero, J., Smith, E., Schevon, C. A., McKhann, G. M., Sheth, S. A., Mehta, A. D. & Mesgarani, N. Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron* **104**, 1195–1209 (2019).
53 Teoh, E. S., Ahmed, F. & Lalor, E. C. Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment. *J. Neurosci.* **42**, 682-691 (2022).
54 Brown, G. J. & Cooke, M. Computational auditory scene analysis. *Computer Speech and Language* **8**, 297-336 (1994).
55 Ellis, D. P. W. *Prediction-driven computational auditory scene analysis* PhD thesis, MIT, (1996).
56 Wang, D. & Brown, G. J. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks* **10**, 684-697 (1999).
57 Barker, J., Cooke, M. P. & Ellis, D. P. W. Decoding speech in the presence of other sources. *Speech Comm.* **45**, 5-25 (2005).
58 Krishnan, L., Elhilali, M. & Shamma, S. A. Segregating complex sound sources through temporal coherence. *PLoS Comp. Biol.* **10**, e1003985, doi:10.1371/journal.pcbi.1003985 (2014).
59 Mlynarski, W. & McDermott, J. H. Ecological origins of perceptual grouping principles in the auditory system. *Proceedings of the National Academy of Sciences* **116**, 25355-25364 (2019).
60 Cusimano, M., Hewitt, L. & McDermott, J. H. Listening with generative models. *Cognition* **253**, 105874 (2024).
61 Boynton, G. M. Attention and visual perception. *Curr. Opin. Neurobiol.* **15**, 465–469 (2005).
62 Wolfe, J. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review* **1**, 202-238 (1994).

63 Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N. & Nuflo, F. Modeling visual attention via selective tuning. *Artificial Intelligence* **78**, 507-545 (1995).
64 Navalpakkam, V. & Itti, L. Search goal tunes visual features optimally. *Neuron* **53**, 605–617 (2007).
65 Eckstein, M. P., Peterson, M. F., Pham, B. T. & Droll, J. A. Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision Research* **49**, 1097-1128 (2009).
66 Whiteley, L. & Sahani, M. Attention in a Bayesian framework. *Frontiers in Human Neuroscience* **6**, 100 (2012).
67 Denison, R. N., Carrasco, M. & Heeger, D. J. A dynamic normalization model of temporal attention. *Nature Human Behaviour* **5**, 1674–1685 (2021).
68 Lindsay, G. W. & Miller, K. D. How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife* **7**, doi:10.7554/eLife.38105 (2018).
69 Konkle, T. & Alvarez, G. Cognitive steering in deep neural networks via long-range modulatory feedback connections. *Advances in Neural Information Processing Systems* **36**, 21613-21634 (2023).
70 Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q. & Kreiman, G. Finding any Waldo with zero-shot invariant and efficient visual search. *Nature Communications* **9**, 3730 (2018).
71 Gupta, S. K., Zhang, M., Wu, C. C., Wolfe, J. & Kreiman, G. in *Advances in Neural Information Processing Systems* Vol. 34 6946-6959 (2021).
72 Srivastava, S., Wang, W. Y. & Eckstein, M. P. Emergent human-like covert attention in feedforward convolutional neural networks. *Current Biology* **34**, 579–593 (2024).
73 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. in *Advances in Neural Information Processing Systems* Vol. 30 (eds I. Guyon *et al.*) (Curren Associates, Inc., 2017).
74 Fritz, J. B., Shamma, S. A., Elhilali, M. & Klein, D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience* **6**, 1216-1223 (2003).
75 Broadbent, D. E. *Perception and Communication*. (Oxford University Press, 1958).
76 Moray, N. Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology* **11**, 56-60 (1959).
77 Treisman, A. M. Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology* **12**, 242-248 (1960).
78 Deutsch, J. A. & Deutsch, D. Attention: Some theoretical considerations. *Psychological Review* **70**, 80-90 (1963).
79 Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782-784 (1985).
80 Lavie, N. Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance* **21**, 451 (1995).
81 Blaser, E., Pylyshyn, Z. W. & Holcombe, A. O. Tracking an object through feature space. *Nature* **408**, 196-199 (2000).
82 Woods, K. J. P. & McDermott, J. H. Attentive tracking of sound sources. *Current Biology* **25**, 2238-2246 (2015).
83 Wostmann, M., Herrmann, B., Maess, B. & Obleser, J. Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences* **113**, 3873-3878 (2016).
84 Stimberg, M., Brette, R. & Goodman, D. F. M. Brian 2, an intuitive and efficient neural simulator. *eLife* **8**, e47314 (2019).

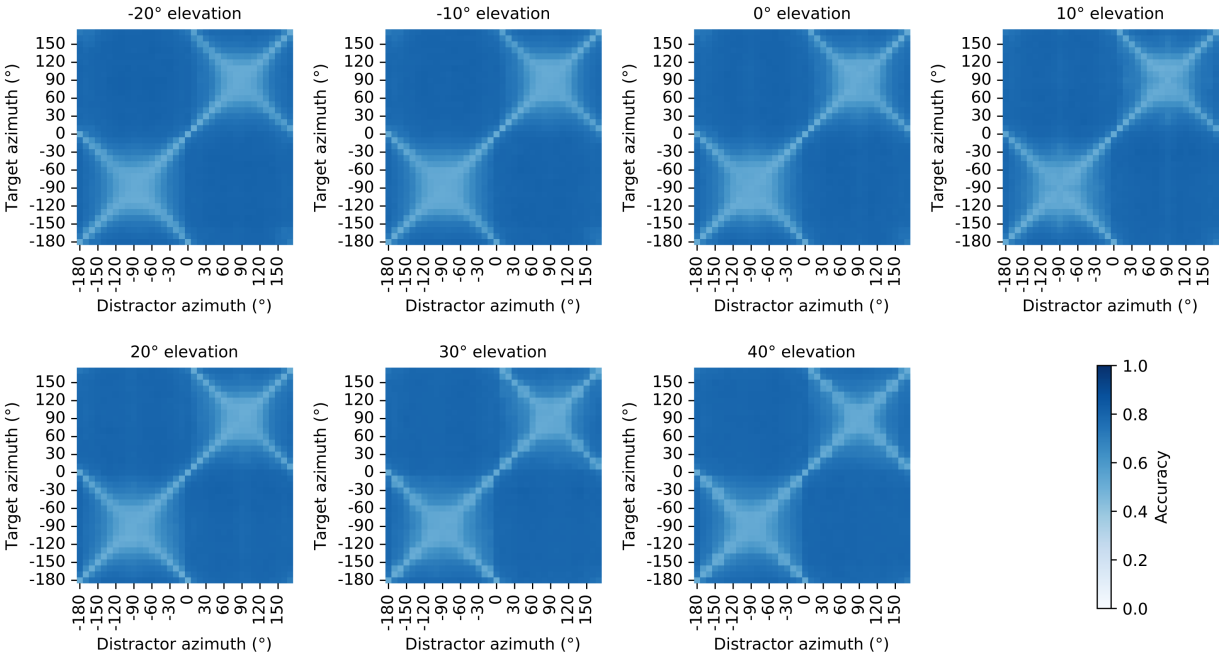
85 Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T. & Rigoll, G. in *Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science* Vol. 12335 (eds A. Karpov & R. Potapova) (Springer, 2020).
86 Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A. & Auli, M. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research* **25**, 1-52 (2024).
87 Best, V., Carlile, S., Jin, C. & van Schaik, A. The role of high frequencies in speech localization. *Journal of the Acoustical Society of America* **118**, 353-363, doi:10.1121/1.1926107 (2005).
88 Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M. & Ritter, M. in *IEEE ICASSP 2017.*
89 Saddler, M. R., Gonzalez, R. & McDermott, J. H. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications* **12**, 7278, doi:10.1038/s41467-021-27366-6 (2021).
90 Feather, J., Leclerc, G., Madry, A. & McDermott, J. H. Model metamers illuminate divergences between biological and artificial neural networks. *Nature Neuroscience* **26**, 2017-2034 (2023).
91 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 Biology*, in press (2023).
92 Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. *arXiv*, 1607.06450 (2016).
93 Ioffe, S. & Szegedy, C. in *International Conference on Machine Learning.* 448-456 (PMLR).
94 Heeger, D. J. Normalization of cell responses in cat striate cortex. *Visual Neuroscience* **9**, 181-197 (1992).
95 Feather, J., Durango, A., Gonzalez, R. & McDermott, J. H. in *Advances in Neural Information Processing Systems (NeurIPS)* (2019).
96 Loshchilov, I. & Hutter, F. in *International Conference on Learning Representations* (2019).
97 Woods, K. J. P. & McDermott, J. Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences* **115**, E3313-E3322 (2018).
98 McWalter, R. & McDermott, J. H. Illusory sound texture reveals multi-second statistical completion in auditory scene analysis. *Nature Communications* **10**, 5096 (2019).
99 McPherson, M. J. & McDermott, J. H. Time-dependent discrimination advantages for harmonic sounds suggest efficient coding for memory. *Proceedings of the National Academy of Sciences* **117**, 32169-32180, doi:10.1101/2020.05.07.082511 (2020).
100 McPherson, M. J., Dolan, S. E., Durango, A., Ossandon, T., Valdés, J., Undurraga, E. A., Jacoby, N., Godoy, R. A. & McDermott, J. H. Perceptual fusion of musical notes by native Amazonians suggests universal representations of musical intervals. *Nature Communications* **11**, 2786 (2020).
101 Traer, J., Norman-Haignere, S. V. & McDermott, J. H. Causal inference in environmental sound recognition. *Cognition* **214**, 104627 (2021).
102 McPherson, M. J., Grace, R. C. & McDermott, J. H. Harmonicity aids hearing in noise. *Attention, Perception, and Psychophysics* **84**, 1016-1042 (2022).
103 Woods, K. J. P., Siegel, M. H., Traer, J. & McDermott, J. H. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics* **79**, 2064-2072 (2017).
104 Köhn, A., Stegen, F. & Baumann, T. in *The Tenth International Conference on Language Resources and Evaluation (LREC'16).* 4644-4647.

105 Kawahara, H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology* **27**, 349-353 (2006).
106 McDermott, J. H., Ellis, D. P. W. & Kawahara, H. in *Proceedings of SAPA-SCALE.* 114-117.
107 McDermott, J. H. & Simoncelli, E. P. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron* **71**, 926-940, doi:10.1016/j.neuron.2011.06.032 (2011).
108 Kidd, G., Best, V. & Mason, C. R. Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *Journal of the Acoustical Society of America* **124**, 3793–3802 (2008).
109 Marrone, N., Mason, C. R. & Kidd, G. The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *Journal of the Acoustical Society of America* **124**, 3064-3075 (2008).
110 Seabold, S. & Perktold, J. in *Proceedings of the 9th Python in Science Conference* (2010).
111 Vallat, R. Pingouin: statistics in Python. *Journal of Open Source Software* **3**, 31, doi:https://doi.org/10.21105/joss.01026 (2018).
112 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & Contributors, S. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261-272, doi:10.1038/s41592-019-0686-2 (2020).
113 Freyman, R. L., Balakrishnan, U. & Helfer, K. S. Spatial release from informational masking in speech recognition. *Journal of the Acoustical Society of America* **109**, 2112-2122 (2001).
114 Arbogast, T. L., Mason, C. R. & Kidd, G. The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America* **112**, 2086-2098 (2002).
115 Levitt, H. & Rabiner, L. R. Binaural release from masking for speech and gain in intelligibility. *Journal of the Acoustical Society of America* **42**, 601-608 (1967).
116 Zurek, P. M. in *Acoustical Factors Affecting Hearing Aid Performance, Second Edition* (eds G.A. Studebaker & I. Hochberg) 255-276 (Allyn and Bacon, 1993).

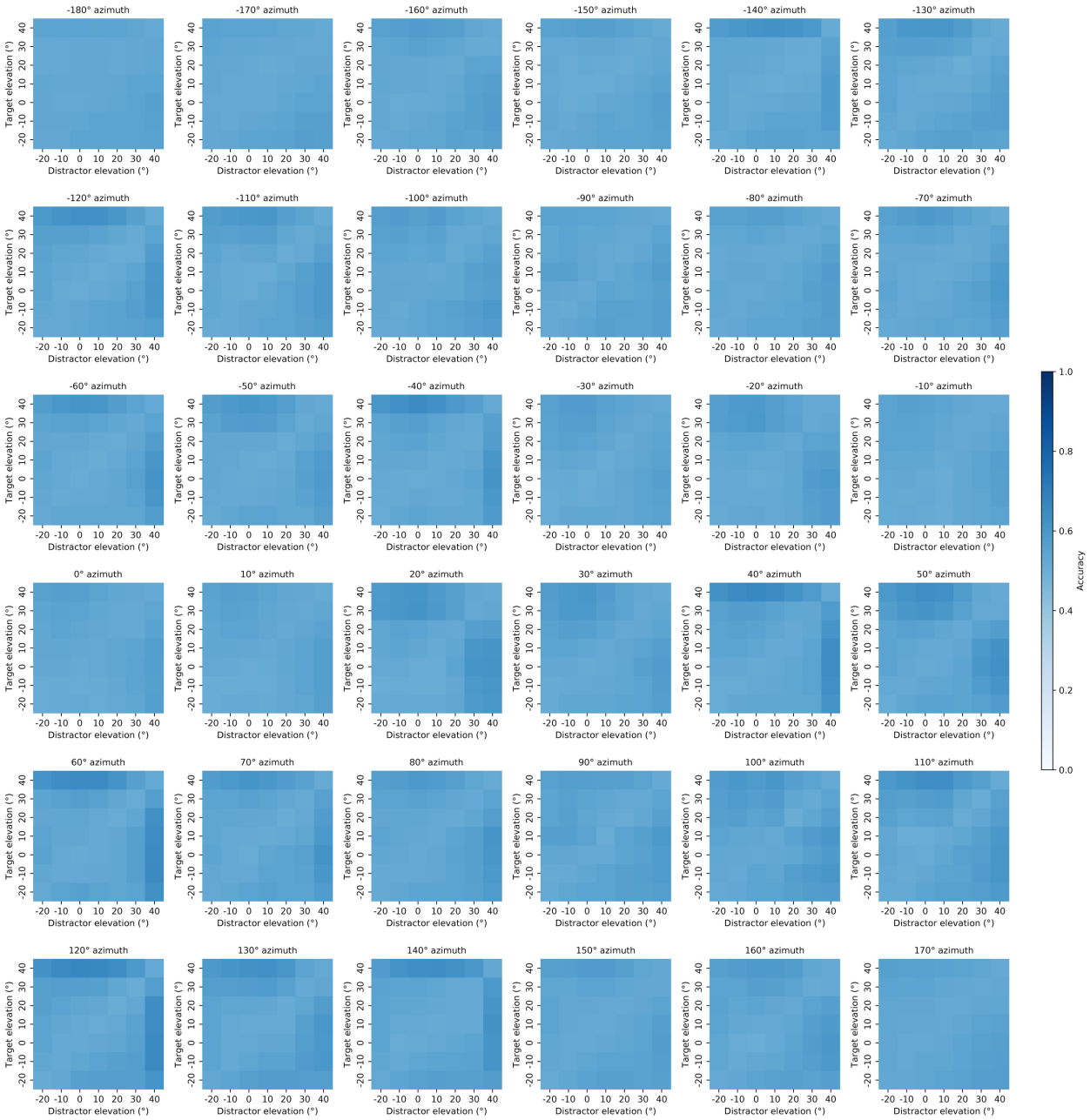
Supplementary Figures and Tables for Griffith, Hess & McDermott



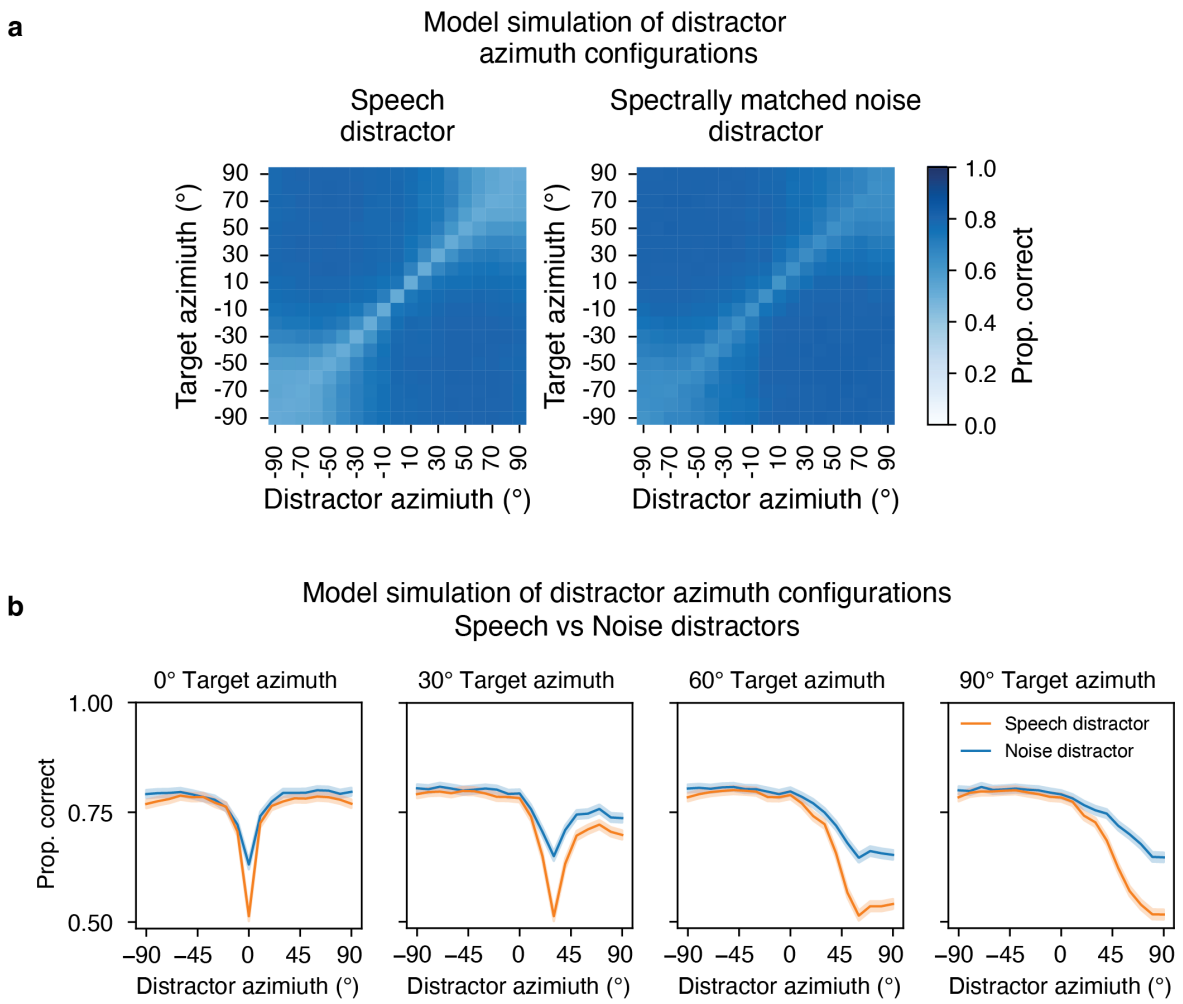
Supplementary Fig. 1 | Performance of all models on Experiment 1. Word recognition as a function of signal-to-noise ratio and distractor type for each model architecture. Panels a-j show results for each individual feature-gain model architecture (main figures show results averaged across these ten architectures, with different conditions plotted separately to elucidate particular effects of interest). Panel k shows results for human participants, replotted from Figure 2. Panels l-n show the baseline, early-only, and late-only architectures which each perform worse with speech-on-speech examples than both humans and the feature-gain models, particularly with 1-distractor (light-pink, gold, green, and blue lines).



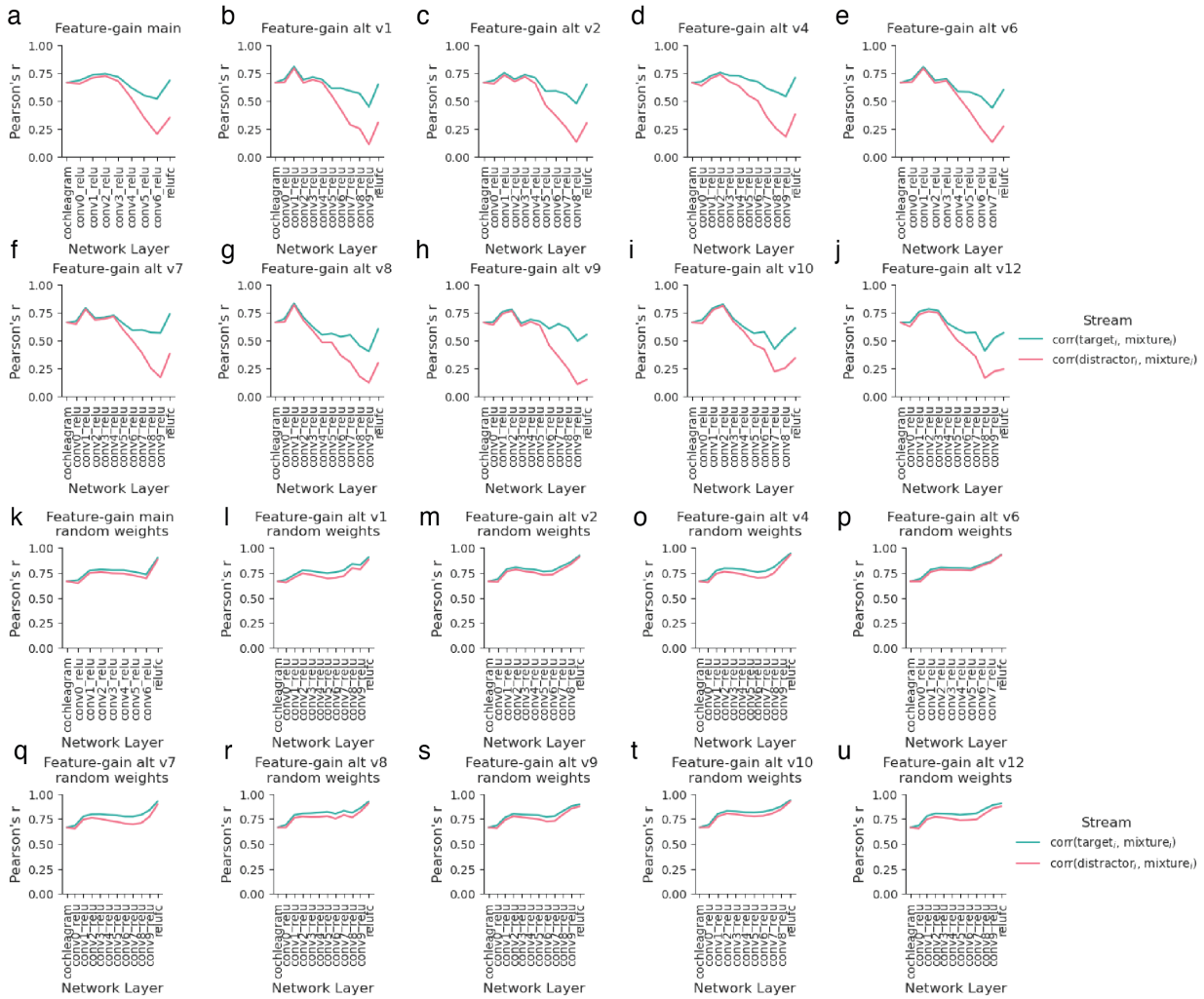
Supplementary Fig. 2 | Model word recognition at all target-distractor azimuth pairings. Word recognition performance for the main feature-gain model (arch_v00 in Supplementary Table 1) as a function of target and distractor positions in azimuth. Each panel plots results for a different elevation, with target azimuth on the vertical axis and distractor azimuth on the horizontal axis (targets and distractors were at the same elevation). The plots differ from those in Figure 4 in plotting results for the full 360 degrees of azimuthal positions (for completeness). Because there was little effect of whether a source was in the front or back hemisphere, the plots in Figure 4 averaged the two hemispheres. The front-back symmetry yields the "X" structures evident in the plots. The broadening of spatial acuity as the target azimuth moved peripherally (i.e. approaching $\pm 90°$ from $0°$) was apparent at all elevations.



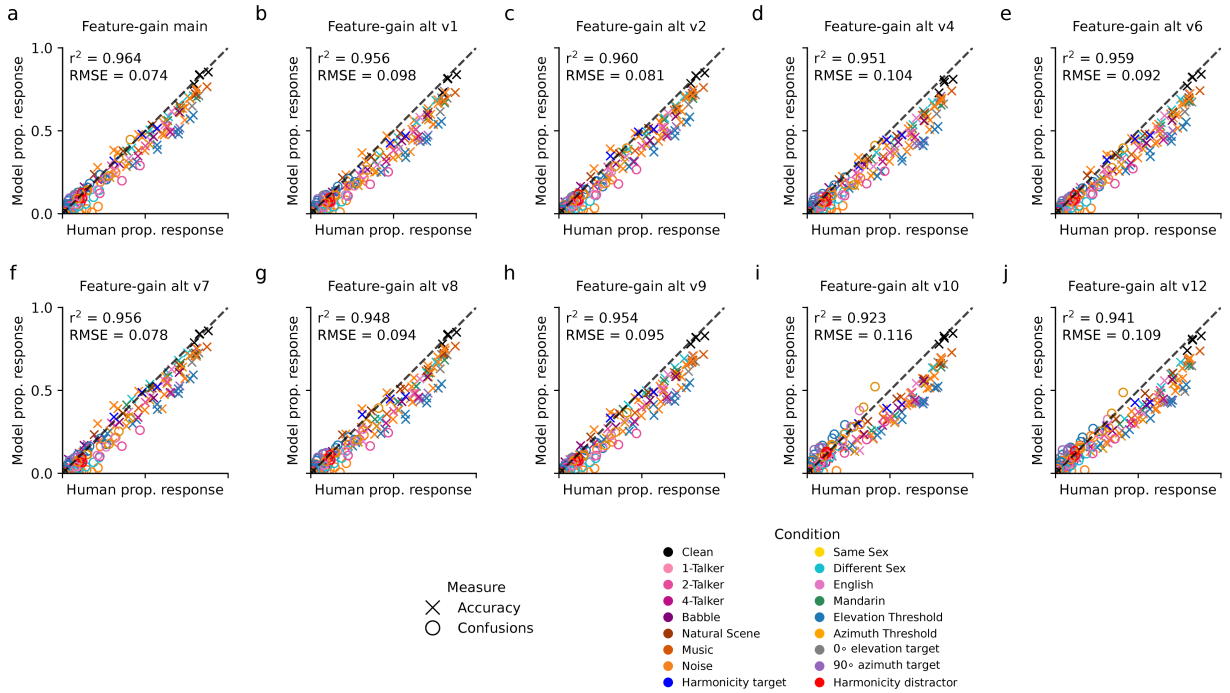
Supplementary Fig. 3 | Model word recognition at all target-distractor elevation pairings. Word recognition performance for the main feature gain model (arch_v00 in Supplementary Table 1) as a function of target and distractor positions in elevation. Each panel plots results for a different azimuth, with target elevation on the vertical axis and distractor elevation on the horizontal axis (targets and distractors were at the same azimuth). Target-distractor offset in elevation produced little benefit regardless of azimuthal position.



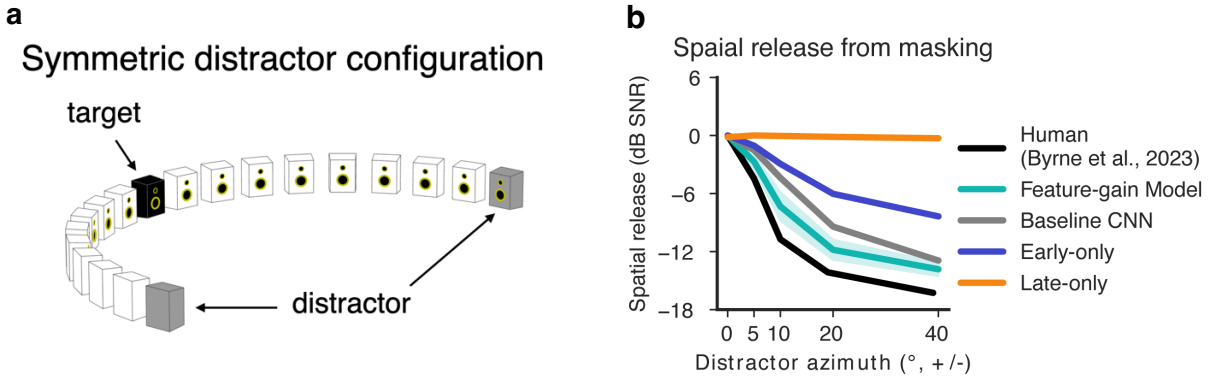
Supplementary Fig. 4 | Effect of target and distractor spatial position on model performance for speech and noise distractors. a. Word recognition performance for the main feature-gain model (arch_v00 in Supplementary Table 1) is plotted as a function of target and distractor positions in azimuth, for both single-talker distractors and speech-shaped noise distractors. As in Figure 4, results are averaged across front and back positions for ease of visualization. **b**. Data from (a) plotted as line graphs for four example target positions, to enable direct comparison of performance for speech and noise distractors. Shaded regions plot confidence intervals (95%) obtained via bootstrap. Performance with noise distractors varies much less with spatial position than performance with speech distractors, as in humans[113,114]. The presumptive explanation is that voice features alone are sufficient to select speech from noise, such that there is less benefit from spatial attention. There is nonetheless some spatial benefit on performance, consistent with prior results in human listeners[114]. This spatial dependence could be explained by a combination of changes in signal-to-noise ratio at the ear (which improves at the ear closer to the target as the distractor is moved away) and binaural effects related to masking[115,116]. However, the larger effect for speech distractors suggests a contribution of spatial attention in this setting. The similar spatial dependence for speech and noise distractors presumably reflects the spatial dependence of the underlying binaural cues, which also cause localization acuity to be worse in the periphery.



Supplementary Fig. 5 | Stage-of-selection analysis for each model architecture. Analysis of Fig. 5, plotted separately for each model architecture. All feature-gain architectures displayed fairly similar patterns of target enhancement at late model stages (panels a-j). Each architecture initialized with random weights did not show the same trend (panels k-u).



Supplementary Fig. 6 | Comparison of human and model performance for each model architecture
Scatter plots of model vs. human performance for each individual feature-gain architecture.



Supplementary Fig. 7 | Spatial release from masking for models lacking architectural constraints. Performance of alternative model architectures (feature-gain, baseline, early-only, and late-only models) as a function of target-distractor separation in azimuth for symmetrically positioned distractors (Experiment 4). Models are compared to human data from Byrne et al. (2024). The y-axis plots spatial release from masking (the decrease in speech reception thresholds compared to the co-located distractor condition) at each tested azimuthal offset. Each of the alternative models displayed less human-like spatial release from masking than the feature-gain model. The late-only model showed no spatial masking release (orange), and the baseline (grey) and early-only model (purple) exhibited degraded masking release compared to the feature-gain model.

Architecture	arch_v00	arch_v01	arch_v02	arch_v04	arch_v06	arch_v07	arch_v08	arch_v09	arch_v10	arch_v12
Operation	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]	Input [40,20000,2]
1	Coch Lnorm	Coch Lnorm	Coch Lnorm	Coch Lnorm	Coch Lnorm	Coch Lnorm	Coch Lnorm	Coch Lnorm	Coch Lnorm	Coch Lnorm
2	Feature-gains 0	Feature-gains 0	Feature-gains 0	Feature-gains 0	Feature-gains 0	Feature-gains 0	Feature-gains 0	Feature-gains 0	Feature-gains 0	Feature-gains 0
3	Lnorm0	Lnorm0	Lnorm0	Lnorm0	Lnorm0	Lnorm0	Lnorm0	Lnorm0	Lnorm0	Lnorm0
4	conv0 [2,34, 32]	conv0 [1,59, 32]	conv0 [2,16, 32]	conv0 [3, 55, 64]	conv0 [1, 22, 32]	conv0 [2, 71, 32]	conv0 [1, 32, 32]	conv0 [2, 11, 32]	conv0 [1, 16, 32]	conv0 [3, 75, 32]
5	Relu0	Relu0	Relu0	Relu0	Relu0	Relu0	Relu0	Relu0	Relu0	Relu0
6	HannPool0 [2,4]	HannPool0 [2,1]	HannPool0 [2,5]	HannPool0 [2,3]	HannPool0 [2,3]	HannPool0 [2,5]	HannPool0 [2,6]	HannPool0 [2,2]	HannPool0 [2,3]	HannPool0 [2,2]
7	Feature-gains 1	Feature-gains 1	Feature-gains 1	Feature-gains 1	Feature-gains 1	Feature-gains 1	Feature-gains 1	Feature-gains 1	Feature-gains 1	Feature-gains 1
8	Lnorm1	Lnorm1	Lnorm1	Lnorm1	Lnorm1	Lnorm1	Lnorm1	Lnorm1	Lnorm1	Lnorm1
9	conv1 [2, 4, 64]	conv1 [2,72, 64]	conv1 [2, 74, 64]	conv1 [2,32,128]	conv1 [1, 34, 64]	conv1 [1, 33, 64]	conv1 [3, 47, 64]	conv1 [3, 60, 64]	conv1 [2, 79, 64]	conv1 [2, 44, 64]
10	Relu1	Relu1	Relu1	Relu1	Relu1	Relu1	Relu1	Relu1	Relu1	Relu1
11	HannPool1 [2,4]	HannPool1 [2,5]	HannPool1 [2,6]	HannPool1 [2,4]	HannPool1 [2,5]	HannPool1 [2,1]	HannPool1 [2,4]	HannPool1 [2,6]	HannPool1 [2,5]	HannPool1 [2,6]
12	Feature-gains 2	Feature-gains 2	Feature-gains 2	Feature-gains 2	Feature-gains 2	Feature-gains 2	Feature-gains 2	Feature-gains 2	Feature-gains 2	Feature-gains 2
13	Lnorm2	Lnorm2	Lnorm2	Lnorm2	Lnorm2	Lnorm2	Lnorm2	Lnorm2	Lnorm2	Lnorm2
14	conv2 [5,5, 256]	conv2 [4,3, 128]	conv2 [5, 4, 128]	conv2 [5, 3, 256]	conv2 [3, 5, 128]	conv2 [3, 6, 128]	conv2 [4, 4, 128]	conv2 [4, 5, 128]	conv2 [5, 4, 128]	conv2 [5, 4, 128]
15	Relu2	Relu2	Relu2	Relu2	Relu2	Relu2	Relu2	Relu2	Relu2	Relu2
16	HannPool2 [1,5]	HannPool2 [1,3]	HannPool2 [1,2]	HannPool2 [1,3]	HannPool2 [1,3]	HannPool2 [1,2]	HannPool2 [1,3]	HannPool2 [1,1]	HannPool2 [1,2]	HannPool2 [1,2]
17	Feature-gains 3	Feature-gains 3	Feature-gains 3	Feature-gains 3	Feature-gains 3	Feature-gains 3	Feature-gains 3	Feature-gains 3	Feature-gains 3	Feature-gains 3
18	Lnorm3	Lnorm3	Lnorm3	Lnorm3	Lnorm3	Lnorm3	Lnorm3	Lnorm3	Lnorm3	Lnorm3
19	conv3 [5,5, 512]	conv3 [6,3, 256]	conv3 [6, 5, 256]	conv3 [5, 4, 512]	conv3 [6, 4, 256]	conv3 [5, 3, 256]	conv3 [4, 6, 256]	conv3 [4, 4, 256]	conv3 [6, 6, 256]	conv3 [4, 6, 256]
20	Relu3	Relu3	Relu3	Relu3	Relu3	Relu3	Relu3	Relu3	Relu3	Relu3
21	HannPool3 [1,4]	HannPool3 [1,3]	HannPool3 [1,2]	HannPool3 [1,1]	HannPool3 [1,3]	HannPool3 [1,2]	HannPool3 [1,3]	HannPool3 [1,3]	HannPool3 [1,2]	HannPool3 [1,3]
22	Feature-gains 4	Feature-gains 4	Feature-gains 4	Feature-gains 4	Feature-gains 4	Feature-gains 4	Feature-gains 4	Feature-gains 4	Feature-gains 4	Feature-gains 4
23	Lnorm4	Lnorm4	Lnorm4	Lnorm4	Lnorm4	Lnorm4	Lnorm4	Lnorm4	Lnorm4	Lnorm4
24	conv4 [6 6, 512]	conv4 [4,4, 512]	conv4 [3, 3, 512]	conv4 [6, 6, 512]	conv4 [3, 4, 512]	conv4 [6, 6, 512]	conv4 [3, 6, 512]	conv4 [3, 4, 512]	conv4 [6, 4, 512]	conv4 [5, 6, 512]
25	Relu4	Relu4	Relu4	Relu4	Relu4	Relu4	Relu4	Relu4	Relu4	Relu4
26	HannPool4 [1,1]	HannPool4 [1,2]	HannPool4 [1,1]	HannPool4 [1,1]	HannPool4 [1,2]	HannPool4 [1,3]	HannPool4 [1,2]	HannPool4 [1,3]	HannPool4 [1,3]	HannPool4 [1,1]
27	Feature-gains 5	Feature-gains 5	Feature-gains 5	Feature-gains 5	Feature-gains 5	Feature-gains 5	Feature-gains 5	Feature-gains 5	Feature-gains 5	Feature-gains 5
28	Lnorm5	Lnorm5	Lnorm5	Lnorm5	Lnorm5	Lnorm5	Lnorm5	Lnorm5	Lnorm5	Lnorm5
29	conv5 [5,5, 512]	conv5 [5 3, 512]	conv5 [4, 5, 512]	conv5 [4, 3, 512]	conv5 [6, 4, 512]	conv5 [3, 5, 512]	conv5 [6, 5, 512]	conv5 [4, 3, 512]	conv5 [3, 5, 512]	conv5 [5, 5, 512]
30	Relu5	Relu5	Relu5	Relu5	Relu5	Relu5	Relu5	Relu5	Relu5	Relu5
31	HannPool5 [1,1]	HannPool5 [1,3]	HannPool5 [1,2]	HannPool5 [1,1]	HannPool5 [1,3]	HannPool5 [1,1]	HannPool5 [1,1]	HannPool5 [1,1]	HannPool5 [1,2]	HannPool5 [1,3]
32	Feature-gains 6	Feature-gains 6	Feature-gains 6	Feature-gains 6	Feature-gains 6	Feature-gains 6	Feature-gains 6	Feature-gains 6	Feature-gains 6	Feature-gains 6
33	Lnorm6	Lnorm6	Lnorm6	Lnorm6	Lnorm6	Lnorm6	Lnorm6	Lnorm6	Lnorm6	Lnorm6
34	conv6 [6 6, 512]	conv6 [5,4, 512]	conv6 [6, 6, 512]	conv6 [5, 3, 512]	conv6 [5, 6, 512]	conv6 [6, 3, 512]	conv6 [5, 4, 512]	conv6 [5, 3, 512]	conv6 [4, 4, 512]	conv6 [3, 3, 512]
35	Relu6	Relu6	Relu6	Relu6	Relu6	Relu6	Relu6	Relu6	Relu6	Relu6
36	HannPool6 [2,4]	HannPool6 [1,2]	HannPool6 [1,3]	HannPool6 [1,3]	HannPool6 [1,2]	HannPool6 [1,2]	HannPool6 [1,2]	HannPool6 [1,3]	HannPool6 [1,2]	HannPool6 [1,2]
37	Feature-gains fc	Feature-gains 7	Feature-gains 7	Feature-gains 7	Feature-gains 7	Feature-gains 7	Feature-gains 7	Feature-gains 7	Feature-gains 7	Feature-gains 7
38	flatten	Lnorm7	Lnorm7	Lnorm7	Lnorm7	Lnorm7	Lnorm7	Lnorm7	Lnorm7	Lnorm7
39	fc [512]	conv7 [3,5, 512]	conv7 [6, 6, 512]	conv7 [6, 4, 512]	conv7 [6, 4, 512]	conv7 [4, 4, 512]	conv7 [4, 4, 512]	conv7 [4, 6, 512]	conv7 [5, 3, 512]	conv7 [6, 6, 512]
40	Relufc	Relu7	Relu7	Relu7	Relu7	Relu7	Relu7	Relu7	Relu7	Relu7
41	dropout	HannPool7 [1,3]	HannPool7 [1,2]	HannPool7 [1,3]	HannPool7 [1,2]	HannPool7 [1,3]	HannPool7 [1,1]	HannPool7 [1,3]	HannPool7 [1,2]	HannPool7 [1,3]
42	fc [800]	Feature-gains 8	Feature-gains 8	Feature-gains 8	Feature-gains fc	Feature-gains 8	Feature-gains 8	Feature-gains 8	Feature-gains 8	Feature-gains 8
43		Lnorm8	Lnorm8	Lnorm8	flatten	Lnorm8	Lnorm8	Lnorm8	Lnorm8	Lnorm8
44		conv8 [6,4, 512]	conv8 [6, 4, 512]	conv8 [6, 6, 512]	fc [512]	conv8 [3, 4, 512]	conv8 [5, 5, 512]	conv8 [3, 3, 512]	conv8 [5, 6, 512]	conv8 [4, 4, 512]
45		Relu8	Relu8	Relu8	Relufc	Relu8	Relu8	Relu8	Relu8	Relu8
46		HannPool8 [1,1]	HannPool8 [1,2]	HannPool8 [1,3]	dropout	HannPool8 [1,3]	HannPool8 [1,2]	HannPool8 [1,2]	HannPool8 [1,2]	HannPool8 [1,2]
47		Feature-gains 9	Feature-gains fc	Feature-gains 9	fc [800]	Feature-gains 9	Feature-gains 9	Feature-gains 9	Feature-gains fc	Feature-gains 9
48		Lnorm9	flatten	Lnorm9		Lnorm9	Lnorm9	Lnorm9	flatten	Lnorm9
49		conv9 [6,3, 512]	fc [512]	conv9 [6, 3, 512]		conv9 [3, 3, 512]	conv9 [5, 4, 512]	conv9 [5, 6, 512]	fc [512]	conv9 [5, 5, 512]
50		Relu9	Relufc	Relu9		Relu9	Relu9	Relu9	Relufc	Relu9
51		HannPool9 [1,2]	dropout	HannPool9 [1,3]		HannPool9 [1,3]	HannPool9 [1,3]	HannPool9 [1,1]	dropout	HannPool9 [1,1]
52		Feature-gains fc	fc [800]	Feature-gains fc		Feature-gains fc	Feature-gains fc	Feature-gains fc	fc [800]	Feature-gains fc
53		flatten		flatten		flatten	flatten	flatten		flatten
54		fc [512]		fc [512]		fc [512]	fc [512]	fc [512]		fc [512]
55		Relufc		Relufc		Relufc	Relufc	Relufc		Relufc
56		dropout		dropout		dropout	dropout	dropout		dropout
57		fc [800]		fc [800]		fc [800]	fc [800]	fc [800]		fc [800]
58										

Supplementary Table 1 I Neural network architectures for feature-gain models. Grey bands indicate stages where feature-gain operations occurred. Legend:

- *Lnorm*: layer normalization operation
- 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
- HannPool $[s_f, s_t]$: Hanning window weighted averaging pooling operation with stride s_f in the frequency dimension and stride s_t in the time dimension
- flatten: reshape operation to map a multidimensional tensor to a vector
- dropout: dropout regularization with 50% dropout rate
- fc $[N]$: fully-connected layer with N units