Human-like feature attention emerges in task-optimized models of the cocktail party problem

Ian M. Griffith (iangriffith@g.harvard.edu)
Program in Speech and Hearing Bioscience and Technology, Harvard University
Boston, MA, 02115, USA

R. Preston Hess (rphess@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
Cambridge, MA, 02139, USA

Josh H. McDermott (jhm@mit.edu)
Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, and Center for Brains, Minds and Machines, Massachusetts Institute of Technology
Cambridge, MA, 02139, USA
Program in Speech and Hearing Bioscience and Technology, Harvard University
Boston, MA, 02115, USA

Abstract

Attention enables communication in settings with multiple talkers, selecting sources of interest based on prior knowledge of their features. Decades of research have left two gaps in our understanding of feature-based attention. First, humans succeed at attentional selection in some conditions but fail in others, for reasons that remain unclear. Second, neurophysiology experiments implicate multiplicative gains in selective attention, but it remains unclear whether such gains are sufficient to account for real-world attention-driven behavior. To address these gaps, we optimized an artificial neural network with stimulus-computable feature-based gains for the task of recognizing a cued talker's speech, using binaural audio input (a "cocktail party" setting). Despite not being fit to match humans, the model replicated human performance across a wide range of real-world conditions, showing signs of selection based both on the voice's timbre and spatial location. The results suggest that human-like attentional strategies emerge as an optimized solution to the cocktail party problem, providing a normative explanation for the limits of human performance in this domain.

Keywords: attention, cocktail party problem, computational modeling, deep neural networks, human behavior

Introduction

Everyday communication requires listeners to attend to speech signals, often in settings with multiple competing talkers. Human attentional selection of one voice among others (the "cocktail party problem"), has been demonstrated in many settings (Culling & Stone, 2017; Kidd & Colburn, 2017) and is known to depend on the features of the individual sound sources – both voice properties such as pitch, and spatial location. However, human attentional selection also has limits. Neurophysiological observations suggest a mechanistic account of attentional selection as feature-based multiplicative gains that enhance the perceptual features of an attended object (Treue & Martínez Trujillo, 1999). Human neuroimaging studies similarly show attentional enhancement of attended sound sources (Mesgarani & Chang, 2012; Puvvada & Simon, 2017). However, it remains unclear whether such multiplicative gains are sufficient to account for real-world attention-based abilities. We tested whether a task-optimized model equipped with multiplicative gains applied to sensory representations would replicate the successes and failures of human attention.

Results

Feature-based attention task. Humans and models reported the middle word in a speech excerpt spoken by a cued talker within a mixture of talkers. The cued talker was indicated by a different excerpt of the cued talker's voice, presented prior to the mixture. Both cue and mixture were presented as binaural audio.

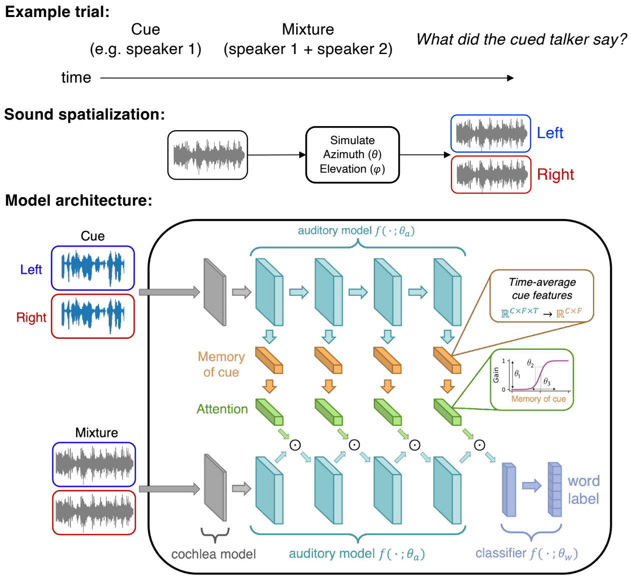

Figure 1: Top: Task is to report word spoken by cued talker in mixture from spatialized audio. Bottom: Diagram of feature-based attention architecture.

Model training. Training examples were constructed by sampling from a set of 48,000 voices, with mixtures composed of a randomly selected target talker excerpt superimposed with between 1 and 5 competing voices or natural sounds, at a randomly sampled signal-to-noise ratio (-10 to 10 dB). Audio was spatially rendered at locations within simulated reverberant rooms using human head-related transfer functions.

Feature-based attention model. We supplemented a convolutional neural network (CNN) model of the auditory system with feature gains derived from the representation of the "cue" sound (Fig. 1). The representation of the cue was derived from the same CNN (blue box in Fig. 1), with features averaged over time to yield a memory representation of the average features of the source. These average cue feature activations were the input to sigmoidal gain functions. Intuitively, the gains should be high for features present in the cue, passing the target talker's features through the auditory system while attenuating other features. The sigmoid parameters were optimized along with the CNN parameters to maximize task performance (by reporting the content of the cued source).

Models attend on par with humans. We measured task performance in 81 human participants, and then simulated the same experiment on the model. Evaluation stimuli were new to participants and the model. As shown in Fig. 2A, the model approximately

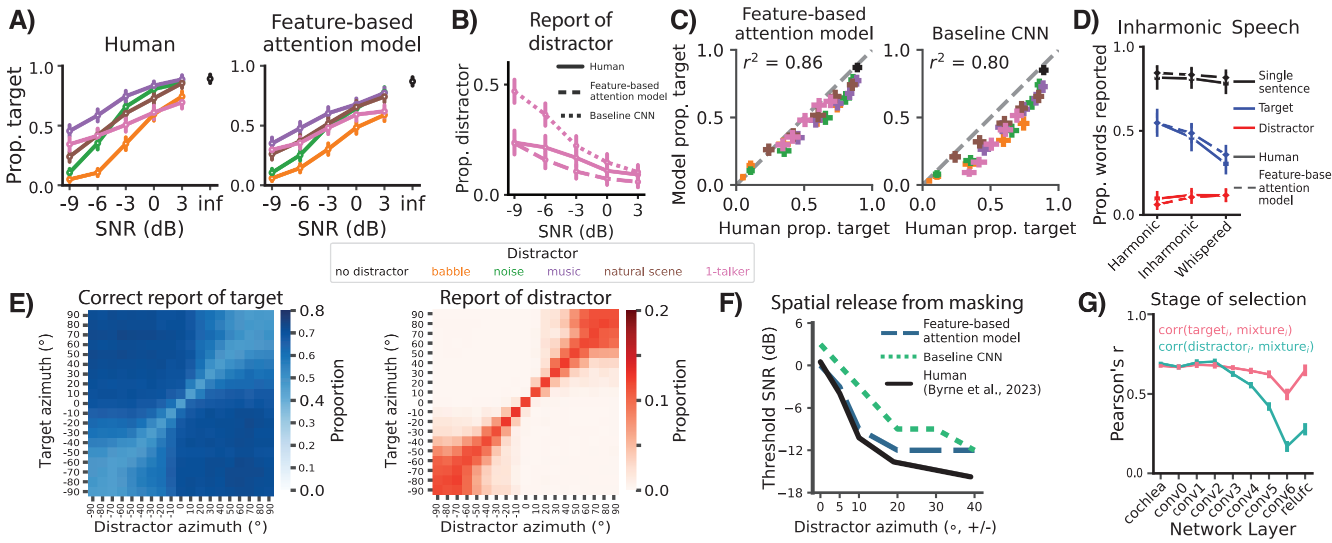

Figure 2: A) Feature-based attention models perform on par with humans. B) Failures of attention (reports of a distractor word) are similarly low for humans and feature-based attention model, but high for Baseline CNN. C) Feature-based attention model explains most of the variance in human performance, better than a Baseline model without attentional gains. D) Feature-based attention model learns human-like reliance on harmonic structure of voices. E) Measuring performance as function of target-distractor proximity shows feature-based attention model learns to attend in space. F) Benefit from spatial separation is similar to humans for models with feature-gains but not for baseline models. G) Difference in target-mixture and distractor-mixture correlations arise only in later model stages, consistent with "late" locus of attention.

replicated both the overall performance of human listeners and the dependence on SNR and distractor type. The model also exhibited failures of attention (reports of words uttered by a distractor talkers) at a similar rate to humans (Fig. 2B), and closely matched the characteristic deficits produced in humans when listening to inharmonic and whispered speech (Fig. 2D).

Models learn spatial tuning in azimuth. Humans benefit from spatial separation between sources ("spatial release from masking"). To test if the models similarly exploited spatial separation, we evaluated performance as a function of target-distractor separation in azimuth. Target word recognition increased and confusion rates decreased as a function of spatial separation for the model (Fig. 2E). To assess whether the model had merely learned to select the ear with the higher SNR, we measured recognition thresholds (the SNR granting 50% of ceiling performance) with distractors placed symmetrically in azimuth (to eliminate "better-ear listening"; Byrne et al., 2023). The model displayed thresholds that varied with spatial separation on par with humans (Fig. 2F).

Models learn late selection. A signature of human auditory attention is the enhancement of the neural representation of a target source at late stages of the auditory hierarchy (Puvvada & Simon, 2017). Attentional selection in the model could in principle occur at any model stage. To assess the locus of attentional selection, at each CNN stage we measured correlations between the activations of target-distractor mixtures and of either the target or distractor alone. Differences between target-mixture and distractor-mixture correlations emerged only at later model stages (Fig. 2G), consistent with human neuroscience evidence.

Baseline models. To investigate the importance of the architectural constraint imposed by the gain functions, we trained a version of the model without this constraint, instead taking the mixture and cue as separate input channels. This Baseline CNN provided a worse match to human behavior, reporting words from "distractor" talkers more frequently (Fig. 2B), and explaining less variance in human performance than the feature-based attention model (Fig. 2C). The Baseline model also had worse absolute performance in spatialized configurations, with higher thresholds than humans (Fig 2F).

Discussion

We provide a framework to model feature-based attention by optimizing a neural network to perform word recognition amid competing talkers. The model provides hypotheses for how attention might be expected to modulate neural responses at different stages of the auditory system, and helps to explain the conditions in which attentional selection is intrinsically difficult for humans.

Acknowledgments

Research supported by National Institutes of Health grant R01 DC017970.

References

Culling, J. F., & Stone, M. A. (2017). Energetic masking and masking release. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party.* New York: Springer international Publishing.

Kidd, G., & Colburn, H. S. (2017). Informational Masking in Speech Recognition. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party.* New York: Springer international Publishing.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Puvvada, K. C., & Simon, J. Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *37*(38), 9189–9196.

Treue, S., & Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579.