Deep neural networks, trained on invariant recognition tasks, struggle to predict hierarchical invariance of speech representations in auditory cortex

Guoyang Liao (guoyang_liao@urmc.rochester.edu)
Biostatistics & Computational Biology, University of Rochester Medical Center
601 Elmwood Ave., Rochester, NY 14642

Dana Boebinger (dana_boebinger@urmc.rochester.edu)
Biostatistics & Computational Biology, University of Rochester Medical Center
601 Elmwood Ave., Rochester, NY 14642

Kirill Nourski (kirill-nourski@uiowa.edu)
Neurosurgery & Iowa Neuroscience Institute, The University of Iowa
200 Hawkins Dr. 1815 JCP, Iowa City IA 52242

Matthew Howard (matthew-howard@uiowa.edu)
Neurosurgery & Iowa Neuroscience Institute, The University of Iowa
200 Hawkins Dr. 1823 JPP, Iowa City IA 52242

Christopher Garcia (christopher-garcia-1@uiowa.edu)
Neurosurgery, The University of Iowa
407 Medical Research Center, Iowa City IA 52242

Thomas Wychowski (thomas_wychowski@urmc.rochester.edu)
Neurology, University of Rochester Medical Center
601 Elmwood Ave., Rochester, NY 14642

Webster Pilcher (webster_pilcher@urmc.rochester.edu)
Neurosurgery, University of Rochester Medical Center
2180 South Clinton Ave., Rochester, NY 14618

Jenelle Feather (jfeather@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute
162 5th Ave, New York, NY 10010

Sam Norman-Haignere (samuel_norman-haignere@urmc.rochester.edu)
Biostatistics & Computational Biology, Neuroscience, Brain & Cognitive Sciences, Biomedical Engineering, University of Rochester Medical Center
601 Elmwood Ave., Rochester, NY 14642

Abstract:

The central computational challenge of speech recognition is that instances of the same class (e.g., word) vary enormously in their acoustics. Traditional auditory models cannot explain "invariant" speech recognition and have difficulty predicting human cortical responses to complex natural stimuli such as speech. Deep neural network (DNN) models trained on challenging invariance tasks such as speech recognition, have shown promise as neural encoding models, but it remains unclear whether they can explain invariant representations of speech in the human auditory cortex. To answer this question, we measured cortical responses to speech with and without acoustic variation using spatiotemporally precise intracranial recordings from neurosurgical patients. We found that representations of speech become increasingly invariant to acoustic variation in non-primary regions, consistent with hierarchical theories of functional organization. We also found DNN models trained on challenging invariance tasks predicted cortical response timecourses to speech better than standard acoustic models, with later network layers better predicting non-primary regions. Yet, all of the tested DNN models had difficulty predicting the hierarchical organization of invariance in the auditory

cortex. These results suggest that the representational invariances learned by current DNN models may not align with those in the auditory cortex.

Keywords: deep neural network; **speech; invariant coding; auditory cortex; intracranial EEG**

Motivation and Background

Speech recognition is computationally challenging because instances of the same class (e.g., a word) vary enormously in their acoustics, due to a myriad of factors (e.g., imperfect sound transmission, reverberation, background noise, variation in speaker articulation). The auditory system is thought to solve this challenge by transforming representations of sound across multiple neuronal processing stages so as to encode sound information in a manner that is robust to acoustic variation (Kell et al., 2018; Sharpee et al., 2011). Understanding and modeling how the auditory system accomplishes invariant coding is central to understanding the neural and computational mechanisms of speech perception (Keshishian et al., 2023), and how they are impaired by neurological disorders (Moore, 1996).

Auditory models have been developed that can effectively predict neural responses early in the auditory pathway (e.g., auditory nerve, inferior colliculus) (Drakopoulos et al., 2021; Nelson & Carney, 2004). These models, however, are unable to explain the impressive speech recognition capacities of human listeners and have difficulty predicting neural responses to complex natural sounds in the human auditory cortex (Kell et al., 2018). Deep neural networks (DNNs) have generated excitement as perceptual and neural models over the past decade because they can be trained to perform challenging invariance tasks such as speech recognition at human levels, and the features learned from these models have been found to be predictive of neural responses throughout sensory cortex (Kell et al., 2018; Kriegeskorte, 2015; Yamins et al., 2014).

Yet, compared with the visual cortex, much less is known about whether task-trained DNNs are capable of serving as effective neural models of the human auditory cortex. There is growing evidence that the human auditory cortex has specialized computational mechanisms for representing speech and music that are not widely present in other species (Landemard et al., 2021), and many human neuroimaging methods, such as fMRI, lack the spatiotemporal precision to track rapidly varying responses to speech in the human auditory cortex (Kell & McDermott, 2019). One prior study found evidence that DNNs can predict intracranially recorded responses to clean speech in the human auditory cortex better than standard acoustic models (Li et al., 2023), but it remains unknown whether these models can explain invariant speech representations in the presence of challenging forms of acoustic variation.

Approach and Results

We measured human cortical responses to speech with and without additional acoustic variation from 13 neurosurgical patients implanted with stereotactic depth electrodes at the University of Rochester Medical Center and the University of Iowa Hospital and Clinics. We tested many different types of acoustic variation: spectral filtering (low-pass, bandpass, or high-pass filter), reverberation (convolution with 12 different naturally recorded impulse responses; Traer et al., 2021), background noise (12 different backgrounds; 10 dB signal-to-noise ratio), and variation in voicing (by replacing the periodic excitation of speech with a noise excitation, simulating whispering). The type of acoustic variation changed every ~4 seconds. We correlated the broadband gamma (70-140 Hz) response timecourse of each electrode to speech with and without acoustic variation to measure the strength of invariance (**Fig 1A**) and divided this correlation by its maximum possible value, quantified as the reliability of each electrode's response across two repetitions of the same stimulus. Electrodes were localized on the reconstructed cortical surface, and we used each electrode's distance to primary auditory cortex (TE1.1) as a measure of its hierarchical position in the auditory cortex (**Fig 1B**) (Norman-Haignere et al., 2022).

We found that the strength of invariance increased substantially with distance from primary auditory cortex (**Fig 1B**) ($p < 0.01$ via hierarchical bootstrapping across subjects and electrodes), consistent with theories of hierarchical functional organization (Kell et al., 2018). To test whether this change in invariance could be predicted from contemporary DNN models, we fit encoding models mapping the activations from each layer of a task-trained DNN onto the response of each electrode (cross-validated, regularized regression). We tested two convolutional neural network models (CochResNet50, CochCNN9) that were explicitly trained to recognize a "foreground" word and speaker in addition to a "background" environmental sound (Tuckute et al., 2022), as well as a self-supervised, Transformer model (HuBERT) with good transfer performance on invariant speech recognition tasks (Hsu et al., 2021). For comparison, we fit encoding models using a standard spectrotemporal model, with strong prediction accuracy in human auditory cortex (Kell et al., 2018). Because our DNNs were acausal, we allowed each model a single global time-lag to align the

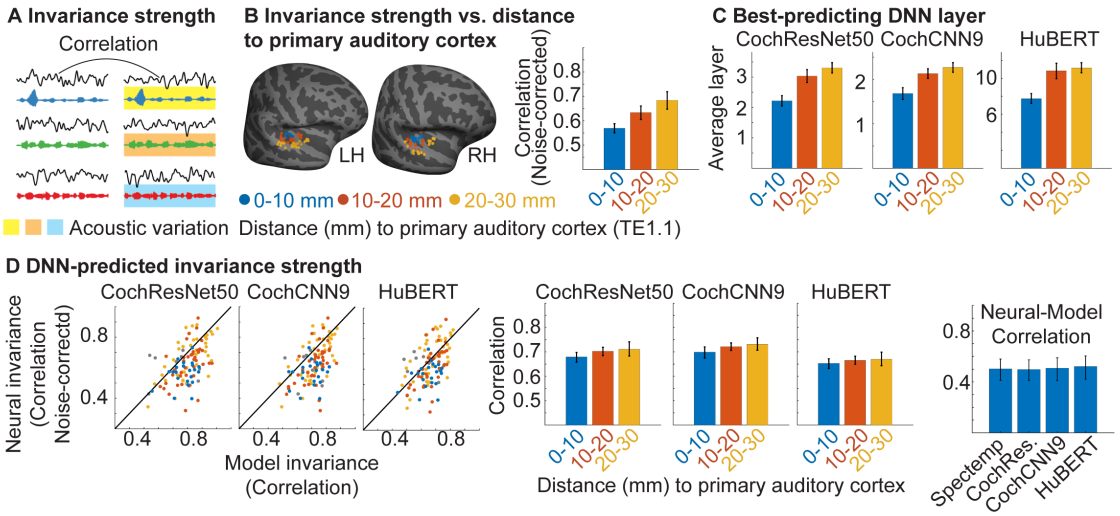

Figure 1. **A** Invariance strength was measured by correlating the neural response to spoken sentences with and without acoustic variation. **B** Average invariance strength (noise-corrected) for electrodes as a function of distance to primary auditory cortex. **C** Best-predicting DNN layer vs. distance to primary auditory cortex. **D** Predicted invariance strength from different DNN models for individual electrodes (left panel) and binned vs. distance to primary auditory cortex (middle panel). Right panel plots the correlation between model and neural invariance for different models.

model features with the neural response (for fairness, the same was done for the spectrotemporal baseline). The time-lag and best-fitting DNN layer were selected in validation data, separate from train and test data.

We first measured the overall model prediction accuracy by correlating the measured and model-predicted response timecourse for each electrode. For all three DNN models, we observed a highly consistent improvement in prediction accuracy relative to the spectrotemporal baseline model ($p < 0.001$ for all DNNs). We also found that all three models replicated hierarchical cortical organization, with later layers better predicting more non-primary regions (**Fig 1C**) ($p < 0.001$ for all DNNs). These findings replicate prior work showing that DNNs trained on challenging speech tasks learn features that are predictive of cortical responses in the human auditory cortex (Kell et al., 2018; Li et al., 2023; Tuckute et al., 2022).

To test whether the DNN models could explain the hierarchical organization of invariance, we correlated the predicted response from each model to speech with and without variation (**Fig 1D**), as was done for our neural data (**Fig 1A**). We found that the predicted invariance was quite similar across the auditory cortical hierarchy (**Fig 1D,** middle panel) ($p > 0.1$ for all models) in contrast with what we observed in our neural data (**Fig 1B**) ($p < 0.05$ for an interaction between the neural data and model predictions) and the DNN models did not perform better than our acoustic baseline in predicting the pattern of invariance across auditory cortex (**Fig 1D**, right panel). This finding suggests that despite capturing some aspects of hierarchical cortical organization (**Fig 1C**), these models have difficulty predicting the hierarchical organization of invariance in the auditory cortex. This failure highlights an important limitation of existing DNN models to explain representational invariances in the auditory cortex, providing a challenge and opportunity for future research aimed at improving cortical encoding models.

Acknowledgments

This study was supported by the National Institutes of Health (NIDCD-R00-DC018051 to S.V.N.-H.).

References

Drakopoulos, F., Baby, D., & Verhulst, S. (2021). A convolutional neural-network framework for modelling auditory sensory cells and synapses. *Communications biology*, *4*(1), 827.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460.

Kell, A. J. E., & McDermott, J. H. (2019). Invariance to background noise as a signature of non-primary auditory cortex. *Nature communications*, *10*(1), 1-11. https://www.nature.com/articles/s41467-019-11710-y

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630-644.

https://www.sciencedirect.com/science/article/pii/S0896627318302502

Keshishian, M., Akkol, S., Herrero, J., Bickel, S., Mehta, A. D., & Mesgarani, N. (2023). Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nature Human Behaviour*, *7*(5), 740-753.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417-446. https://doi.org/10.1146/annurev-vision-082114-035447

Landemard, A., Bimbard, C., Demene, C., Shamma, S., Norman-Haignere, S., & Boubenec, Y. (2021). Distinct higher-order representations of natural sounds in human and ferret auditory cortex. *Elife*, *10*. https://doi.org/10.7554/eLife.65566

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., & Chang, E. F. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, *26*(12), 2213-2225.

Moore, B. C. J. (1996). Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear and Hearing*, *17*(2), 133-161. http://journals.lww.com/ear-hearing/Abstract/1996/04000/Perceptual_Consequences_of_Cochlear_Hearing_Loss.7.aspx

Nelson, P. C., & Carney, L. H. (2004). A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. *The Journal of the Acoustical Society of America*, *116*(4), 2173-2186. https://doi.org/10.1121/1.1784442

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*, 455-469. https://doi.org/https://doi.org/10.1038/s41562-021-01261-y

Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology*, *21*(5), 761-767. https://doi.org/10.1016/j.conb.2011.05.027 (Networks, circuits and computation)

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2022). Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence. *BioRxiv*.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624. http://www.pnas.org/content/111/23/8619.short http://www.pnas.org/content/111/23/8619.long