Training on Ecologically Relevant Tasks Improves Alignment Between Artificial Neural Network and Human Similarity Judgments

Aidan Seidle[1], Jenelle Feather[2]*, Malinda J. McPherson-McNato[1]*
[1]Department of Speech, Language and Hearing Sciences, Purdue University
[2] Center for Computational Neuroscience, Flatiron Institute

Abstract

Artificial neural networks (ANNs) have emerged as leading models for predicting human behavior and neural data. While these models have been extensively studied in the visual domain, their effectiveness in modeling audition is comparatively underexplored. Recent work has found that some ANNs can predict aspects of auditory cortical processing, however it is not clear whether these models capture task-invariant representations of sounds. Here, we used human judgments of similarity as a benchmark for the generalization of auditory model representations. We hypothesized that similarity scores computed from models that best predicted neural activation patterns would strongly correlate with human similarity judgments. We compared human similarity judgments of pairs of sounds to cosine similarity calculated from different layers of seventeen ANNs, as well as a basic spectrotemporal model. The ANNs exhibited a wide range of variability in their correlations with human similarity judgments, and the best models were those trained on multiple tasks. Although there was a significant correlation between the ability of a model to predict fMRI data and the alignment with human similarity measurements, some models showed diverging values. This result suggests that separate criteria for correspondence to human behavior, neural data, and higher-level psychological processes may be necessary.

Keywords: similarity; ANNs; auditory

Introduction

State-of-the-art ANNs are generally trained using classification tasks. The extent to which such ANNs a) recapitulate the steps of neural processing and therefore can be used as biological models, or b) learn flexible representations of stimuli and are therefore useful as models of higher-level cognition, are open issues.

Much of the work examining the biological plausibility and generality of ANN stimulus representations has occurred in the domain of vision (Cichy et al., 2016; Jang, McCormack, & Tong, 2021; Jozwik et al., 2017; Peterson, Abbot, & Griffiths, 2018; Yamins et al., 2014). Recent work in audition demonstrated that many auditory models predict fMRI activity evoked by natural sounds, and there exists a correspondence between model layers and cortical regions (Kell et al., 2018; Tuckute & Feather et al., 2023). Studies of auditory ANNs have also shown an alignment between human behavioral judgments and the behavior of deep neural networks. However, comparisons are often performed for specific domains (Francl & McDermott, 2022; Kell et al., 2018; Li et al., 2023; Saddler, Gonzalez, & McDermott, 2021), and rely on explicit categorization decisions. The extent to which auditory ANNs replicate human representational geometry for a broad set of sounds is yet unclear.

Here, we measure human pair-wise similarity ratings for natural sounds and compare these to pair-wise cosine similarity values computed from ANN representations. We investigate a diverse set of auditory ANNs that previous work evaluated as encoding models for fMRI responses. Using sounds from outside the models' training sets, we collected Likert scale similarity ratings for stimuli pairs from humans. We correlated these with cosine similarity measurements from the embedding spaces in the final layers of the ANNs. We find that models vary widely in their ability to predict human similarity judgments and propose this as a possible benchmark for future auditory models.

Methods

Audio dataset. We compiled 706 two-second sound clips, including speech, music, impacts, textures (McDermott & Simoncelli, 2011), and other everyday sounds (Norman-Haignere et al., 2015; Traer et al., 2021). We sought to make this dataset representative of sounds listeners might regularly hear, and make it distinct from model training sets.

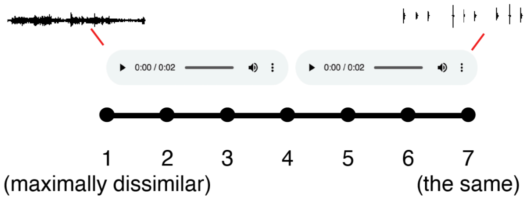

Figure 1. Example human similarity rating task trial with sample sound waves

Human similarity ratings. Fifty participants heard two sounds and were asked to rate how similar they thought the sounds were on a 1-7 point Likert scale (Fig. 1). Participants rated a set of 180 sound pairs, including 20 identical pairs (intended to be catch-trials). Non-idential pairs were randomly selected from 8 bins linearly spanning the cosine similarity scores of the CochResNet50-MultiTask activations to equally represent the entire range of model representations. Two participants who did not correctly identify the majority of the identical catch pairs were excluded. Likert ratings were scaled to 0-1 for plotting. The average split-half reliability of the human data (bootstrapped 10,000 times) was ρ=.93, p<.001.

Model similarity scores. We evaluated the models included in Tuckute & Feather et al., 2023. We passed all 706 sounds through the models, and collected activation vectors from each model layer. Cosine similarity was computed for all pairs of sounds, but to assess the models we only compared cosine similarities for the human-rated pairs (excluding catch trials). We computed Spearman's ρ between human data and cosine similarity from the last 3 layers of each model. We plot the highest of those 3 correlations (Fig. 2). For each model, these similarity scores were compared with the voxelwise prediction values of fMRI data collected on a set of 165 natural sounds (N=8) (data from Norman-Haignere et al., 2015, voxelwise prediction scores from Tuckute & Feather et al., 2023).

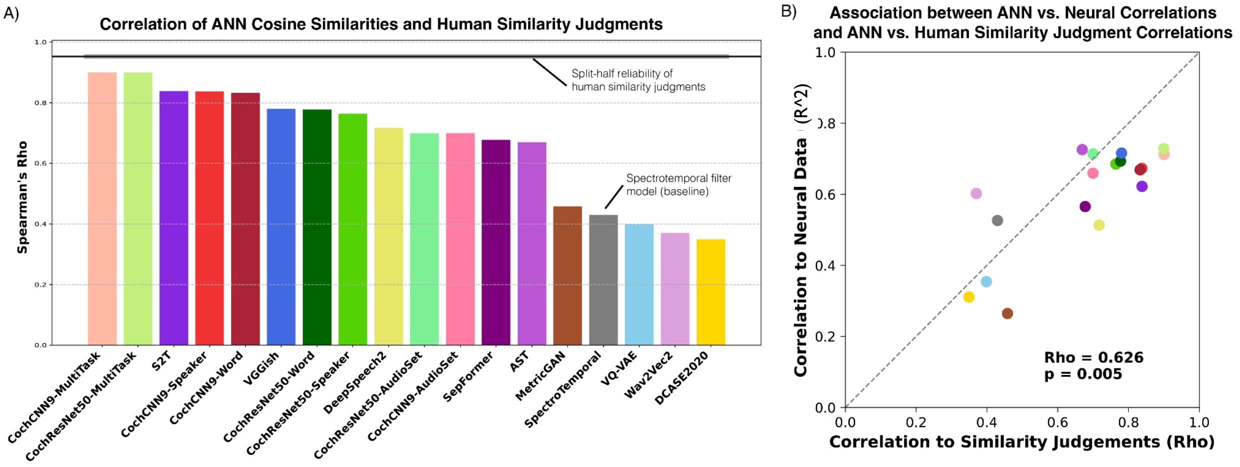

Figure 2. A) Correlation between cosine similarity computed from models and human pairwise similarity judgments. Each bar corresponds to a trained model. All correlations were significant at a p<.001 level. B) Scatter plot of variance explained for neural activation vs. correlation between human and model similarity correlations. The colors of the dots correspond with the bars from Fig. 2A.

Results and Discussion

Overall, models produced similarity measures that ranged widely in their correlations with human similarity judgments (Fig. 2A). Models trained on multiple tasks had the strongest correlations with human data, almost reaching the ceiling of human reliability. Others, however, did not perform as well as a baseline spectrotemporal model.

While we initially predicted that models whose activations were highly correlated with human neural data would also predict human similarity judgments, this did not fully appear to be the case. The variance explained for neural activation was moderately associated with model vs. human similarity correlations (rho=.63, p=.005, Fig. 2B). These results suggest that distinct criteria and optimization may be needed to develop ANNs that predict neural data, behavior, and higher-level psychological phenomena.

References

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep. 2016.

Francl, A., & McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature human behaviour*, *6*(1), 111-133.

Jang H, McCormack D, Tong F. Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. Summerfield C, editor. PLOS Biol. 2021 Dec 9;19(12):e3001418.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, *8*, 1726.

Kell, A.J., Yamins, D.L., 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.

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, 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.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, *71*(5), 926-940.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *neuron*, *88*(6), 1281-1296.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648-2669.

Saddler, M. R., Gonzalez, R., & McDermott, J. H. (2021). Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature communications*, *12*(1), 7278.

Traer, J., Norman-Haignere, S. V., & McDermott, J. H. (2021). Causal inference in environmental sound recognition. *Cognition*, *214*, 104627.

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2023). Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, *21*(12), e3002366.

Yamins D.L., Hong H., Cadieu C.F., Solomon E.A., Seibert D., DiCarlo J.J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci. 2014 Jun.