Discovering the perceptual space of natural sounds from similarity judgments

Jarrod M. Hicks[1,2,3], Bryan J. Medina[1,2,3] & Josh H. McDermott[1,2,3,4]

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

Introduction

- Understanding multidimensional mental representations has been a longstanding challenge in perceptual & cognitive science
- We explored whether representations derived from machine learning could account for human similarity judgments in the domain of audition
- We collected a dataset of human similarity judgments for sound textures and asked:
  1. **How well can representations from contemporary auditory models predict these human judgments?**
  2. **How many dimensions are needed to account for this perceptual space?**

Sound similarity experiment

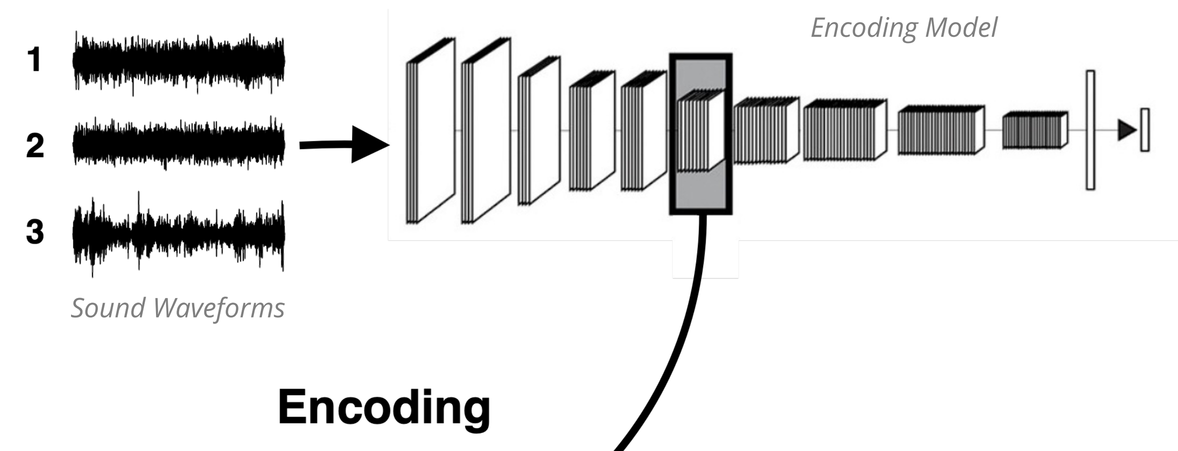
- Triplet odd-one-out task (Hebart et al. 2020)
- On each trial, participants (N=213) listened to three sounds and chose the odd-one-out
- Implicitly indicates which pair of sounds was the most similar within the triplet
- Stimuli: 1,080 unique two-second excerpts of natural sound textures drawn from YouTube soundtracks (AudioSet)
- We collected judgments for 38,332 triplets

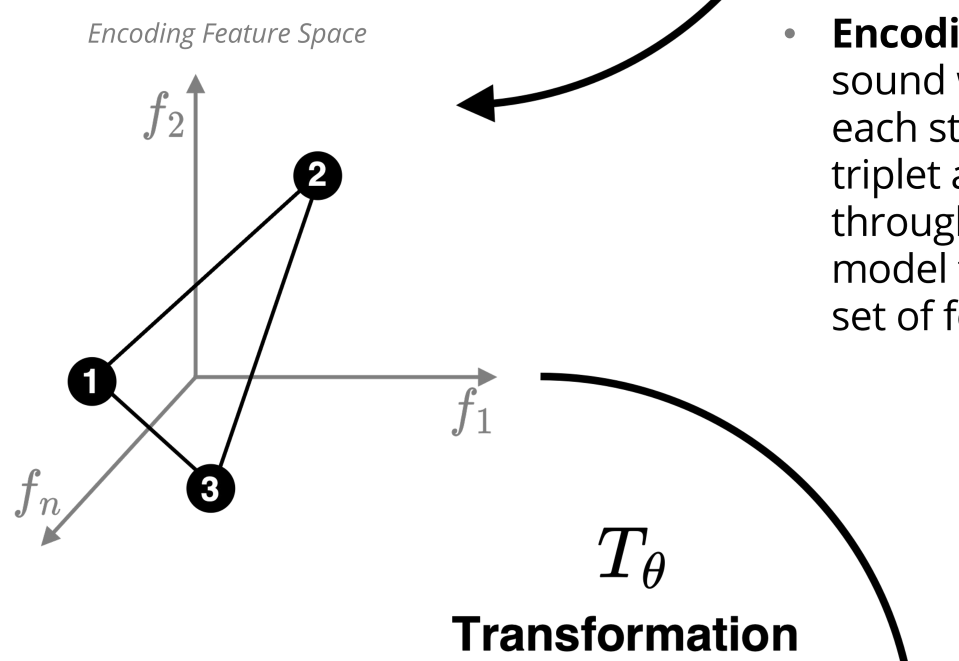
Noise ceiling

- We collected judgments from 20 participants for 180 randomly chosen triplets
- Measured average consistency of choices for each triplet across participants, yielding a noise ceiling of 68.47%
- Provides an upper bound on best possible performance achievable by any model given the variability across participants

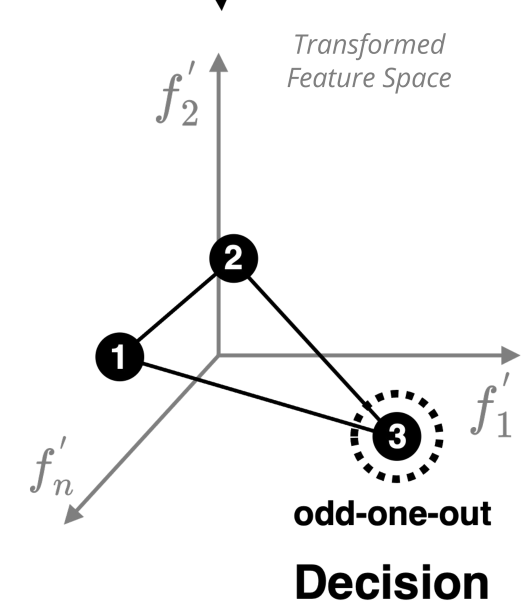
Supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374

Similarity modeling framework



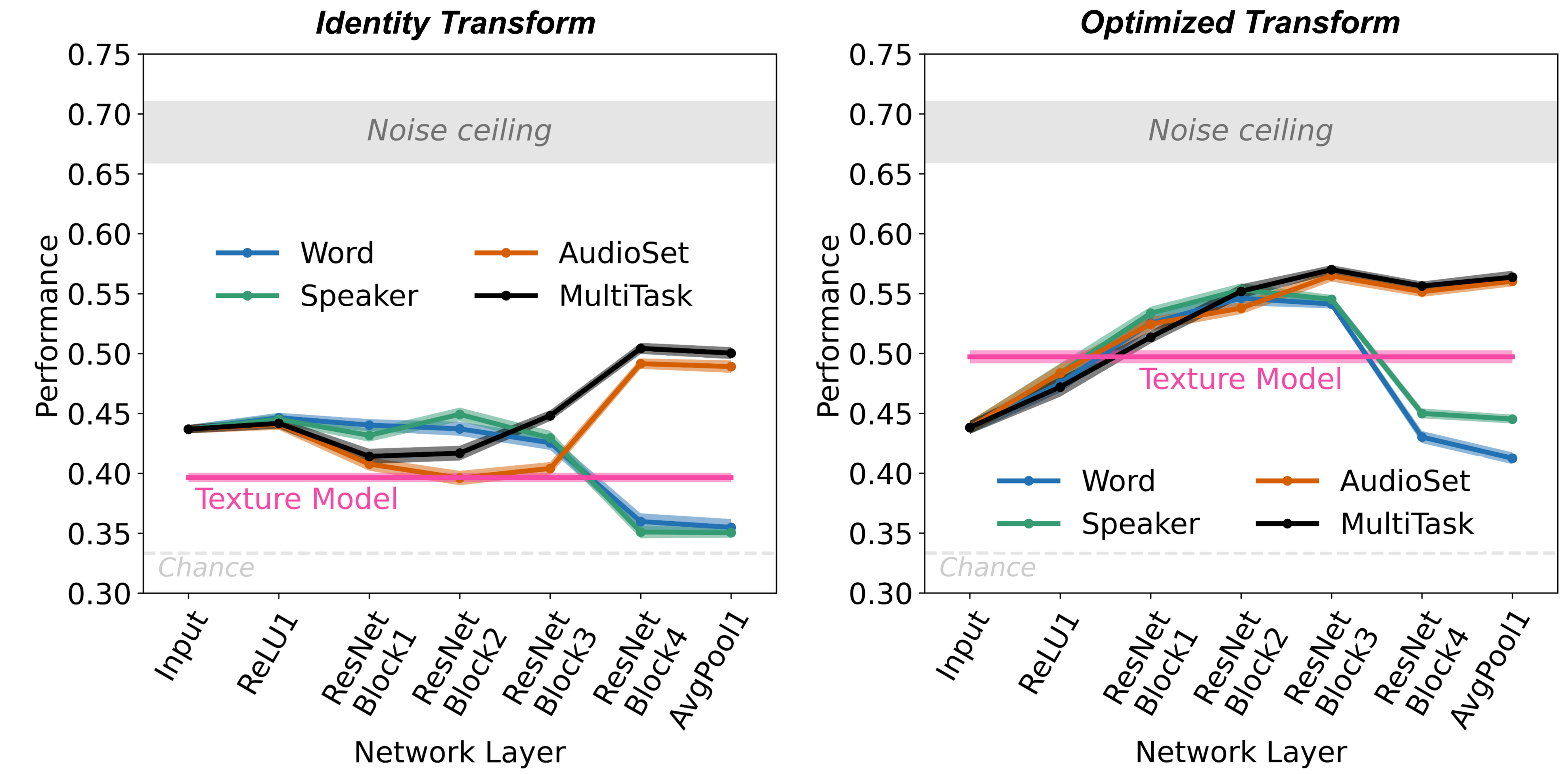


- **Encoding** stage: the sound waveforms for each stimulus within a triplet are passed through an encoding model to generate a set of feature vectors
- **Transformation** stage: feature vectors from the encoding stage are transformed into a new feature space where the odd-one-out decision can be made
- **Decision** stage: the distance between all pairs of stimuli within a triplet is computed in the transformed feature space and the odd-one-out is selected as the stimulus not contained in the minimum-distance pair



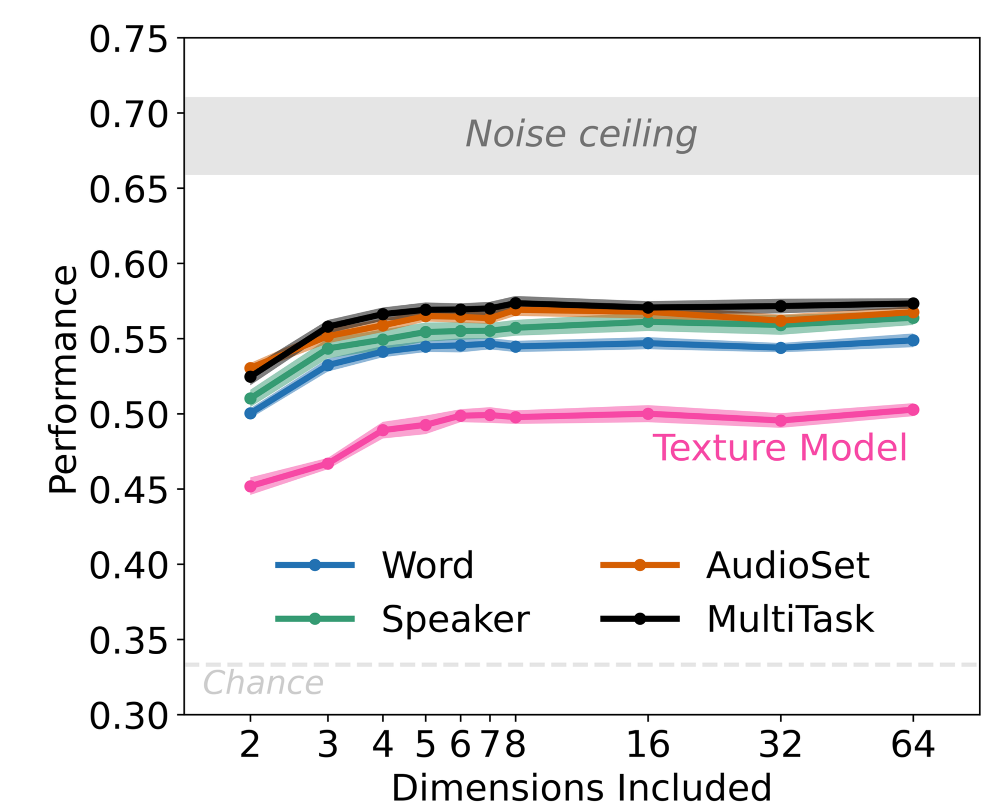
Model performance

- Learned linear transformations were critical to predicting judgments on held-out sounds:
  - an identity transform yielded average performance of 43%
  - the learned transforms yielded performance of 51%, averaged across all stages of all models
- The best-performing stages of the trained neural networks also substantially outperformed the texture model
- Strong dependence on task neural networks were trained on: late stages of models trained to recognize words and speakers produced worse predictions than late stages of models trained on the AudioSet environmental sound recognition task
  - Plausibly explained by these tasks requiring model to be invariant to background noise, perhaps by throwing out information related to textures
- Sizeable gap remained between the best model performance and the noise ceiling



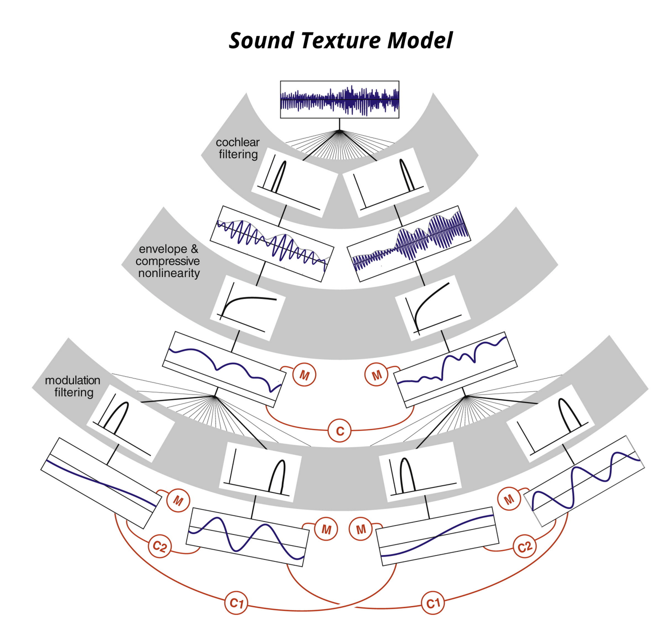
Low-dimensional projections

- For each model, we used the representation from the best-performing stage and learned a linear projection to a low-dimensional feature space, varying the number of dimensions included
- All models reached their peak performance with a surprisingly low number of dimensions
- This result is surprising since many dimensions are needed to synthesize perceptually realistic textures (Feather & McDermott 2018)
- Raises the possibility that similarity judgments tap into a representation that is impoverished relative to that used for discrimination tasks or realism judgments

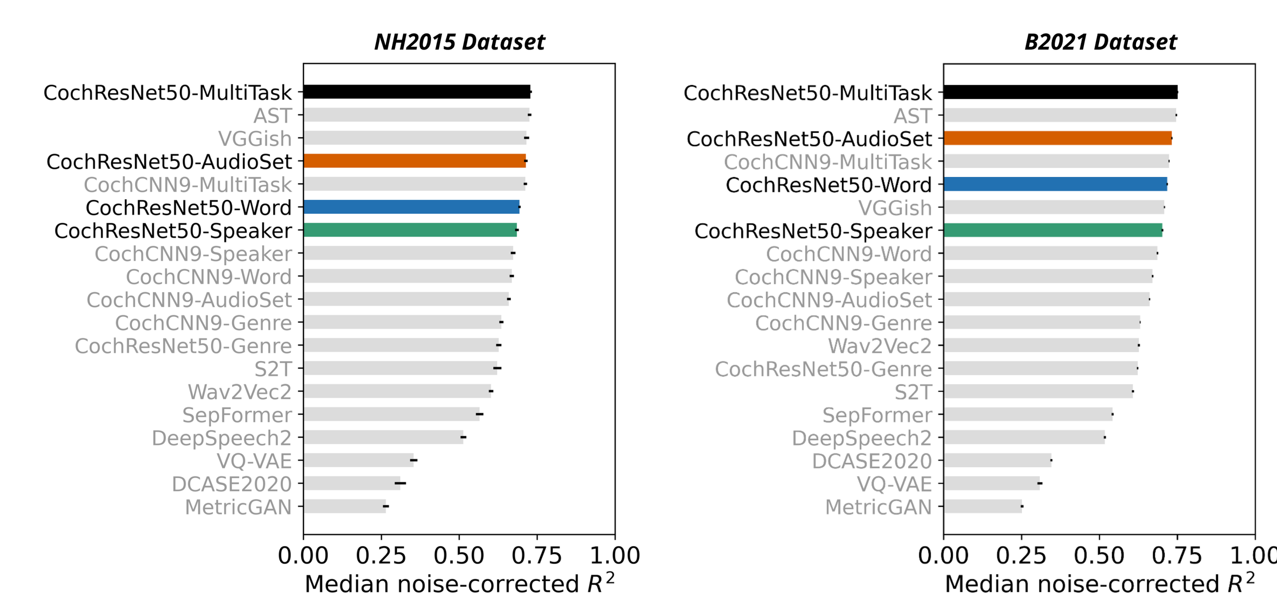


Encoding models

- We evaluated representations from two distinct classes of encoding models:
  1. standard sound ***texture model*** (McDermott & Simoncelli 2011)
  2. convolutional neural network models previously shown to effectively predict neural responses to natural sounds (Tuckute et al. 2023)
- Networks were trained to perform either word recognition ("***Word***"), speaker identification ("***Speaker***"), or background sound recognition ("***AudioSet***") tasks individually, or to perform all three tasks simultaneously ("***MultiTask***")
- We found little effect of network architecture and thus present results for a single ResNet50 architecture with a cochleagram front end ("CochResNet50")



McDermott & Simoncelli 2011



Adapted from Tuckute et al. 2023

Conclusions

- We collected similarity judgments for natural sounds and assessed how well representations from auditory models could predict these judgments
- Linear transformations of model representations substantially improved performance, but a sizeable gap remained between the best model's performance and the noise ceiling, indicating that current models fail to fully capture human sound similarity
- Peak model performance could be achieved with surprisingly low-dimensional representations
- Future work is needed to understand what aspects of perception these dimensions capture and what additional dimensions must be added to improve human-model alignment and close the gap with the noise ceiling