Discovering the perceptual space of natural sounds from similarity judgments

Jarrod M. Hicks (jmhicks@mit.edu)
Department of Brain and Cognitive Sciences, MIT
McGovern Institute for Brain Research, MIT
Center for Brains, Minds and Machines, MIT
Cambridge, MA, 02139, USA

Bryan J. Medina (bjmedina@mit.edu)
Department of Brain and Cognitive Sciences, MIT
McGovern Institute for Brain Research, MIT
Center for Brains, Minds and Machines, MIT
Cambridge, MA, 02139, USA

Josh H. McDermott (jhm@mit.edu)
Department of Brain and Cognitive Sciences, MIT
McGovern Institute for Brain Research, MIT
Center for Brains, Minds and Machines, MIT
Cambridge, MA, 02139, USA

Program in Speech and Hearing Bioscience and Technology, Harvard
Boston, MA, 02115, USA

Abstract

Perceptual similarity is critical to many aspects of perception and cognition, but is poorly characterized for realistic stimuli. We examined the perceptual space of natural sounds using a similarity judgment task applied to large numbers of natural sound textures. Participants judged the similarity of sound textures using an odd-one-out task. We then fit a linear transform to best predict human similarity judgments from a set of candidate representations taken from contemporary auditory models (trained convolutional neural networks or a standard auditory texture model). We found that the learned linear transformations were critical to predicting the human judgments, and that intermediate-to-late stages of the trained neural networks yielded the highest prediction accuracy of human judgments. Surprisingly, only a few dimensions were required to reach peak prediction accuracy. This result suggests that when comparing randomly chosen natural sounds, human similarity is dominated by a small number of dimensions. This general result could constrain memory errors, category formation, and other cognitive phenomena that are dependent on similarity.

Keywords: sound similarity; auditory textures; computational modeling; deep neural networks; human behavior

Introduction

In perceptual and cognitive science, understanding multidimensional mental representations has been a longstanding challenge (Roads & Love 2024). One bottleneck has been the lack of stimulus computable models that can operate on real-world stimuli to produce representations useful for downstream tasks. This has changed in the current era of deep learning where neural networks regularly achieve human-level performance on many real-world tasks and have become leading candidate models in audition and vision. (Kell & McDermott 2019). Here 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, asking 1) how well representations from contemporary auditory models (trained convolutional neural networks or a standard auditory texture model) can predict these human judgments, and 2) how many dimensions are needed to account for this perceptual space.

Methods

Sound similarity experiment. We conducted a sound similarity experiment in which participants performed a 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 indicating which pair of sounds was the most similar within the triplet. Stimuli consisted of 1,080 unique two-second excerpts of natural sound textures drawn from a large set of YouTube soundtracks (AudioSet) (Gemmeke et al. 2017). In total, we collected judgments for 38,332 triplets.

Similarity modeling. To model human judgments, we used a stimulus-computable similarity modeling framework (Fig. 1) consisting of three stages.

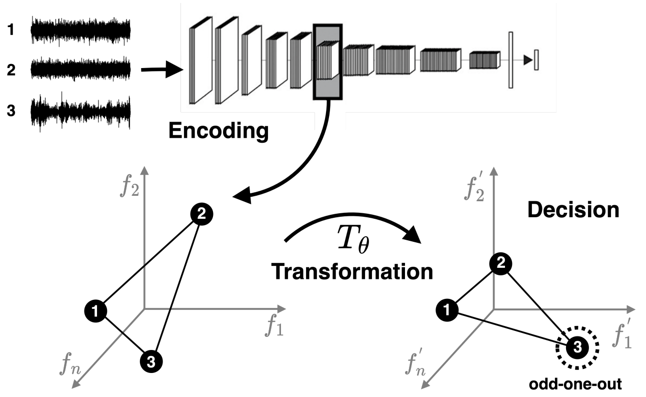

Figure 1: Similarity modeling framework

In the **encoding** stage, the sound waveforms for each stimulus within a triplet are passed through an encoding model to generate a set of feature vectors. The modeling framework is agnostic to the form of this encoding model so long as it takes a sound waveform as input and generates a feature vector.

In the **transformation** stage, feature vectors from the encoding stage are transformed into a new feature space where the odd-one-out decision can be made. While this transformation could be arbitrarily complex (e.g., a neural network), we consider only linear transformations in the present work.

In the **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. We found qualitatively similar results using both cosine and Euclidean distance and thus report results averaged over both distance metrics.

To optimize model parameters, we applied a softmin to the pairwise distances and quantified the error between model choices and human judgments using a cross-entropy loss, then updated parameters using stochastic gradient descent to minimize the loss. While this framework allows for model parameters to be optimized at any model stage, we used frozen encoding and decision stages and only optimized parameters of the transformation stage.

To provide an upper bound on the best possible performance achievable given the variability across participants (i.e., the noise ceiling), we additionally collected 20 judgments for 180 randomly chosen triplets and measured the average consistency of choices for each triplet across participants. This yielded a noise ceiling of 68.47%.

Encoding models. We evaluated representations from two distinct classes of encoding models: a standard sound texture model (McDermott & Simoncelli 2011) and 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, speaker identification, or background sound ("AudioSet") recognition tasks individually, or to perform all three tasks simultaneously ("MultiTask"). As in previous work, we found little effect of network architecture and thus present results for a single ResNet50 architecture with a cochleagram front end ("CochResNet50").

Results

Model performance. The learned linear transformations were critical to predicting judgments on held-out sounds: an identity transform yielded average performance of 43%, whereas the learned transforms yielded 51%, averaged across all stages of all models (Fig. 2A). The best performing stages of the trained neural networks also substantially outperformed the texture model. We also observed a strong dependence on the task the neural networks were trained on. In particular, the late stages of the models trained to recognize words and speakers produced worse predictions than the late stages of the models trained on the AudioSet environmental sound recognition task. This latter result is plausibly explained by the fact that these tasks require the model to be invariant to background noise, which might be achieved by throwing out information related to textures. Nonetheless, a 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 from the previous analysis and learned a linear projection to a low-dimensional feature space, varying the number of dimensions included. We found that all models reached their peak performance with a surprisingly low number of dimensions (Fig. 2B). To assess whether this low-dimensionality was related to the amount of human judgments used for training, we re-ran the analysis using only half of the human data and found that the results were highly similar to that using the full data (r=0.99), and only slightly worse in absolute terms. This result is surprising in light of findings that large numbers of dimensions are needed to synthesize perceptually realistic textures (Feather & McDermott 2018) and raises the possibility that similarity judgments tap into a representation that is impoverished relative to that used for discrimination or realism judgments.

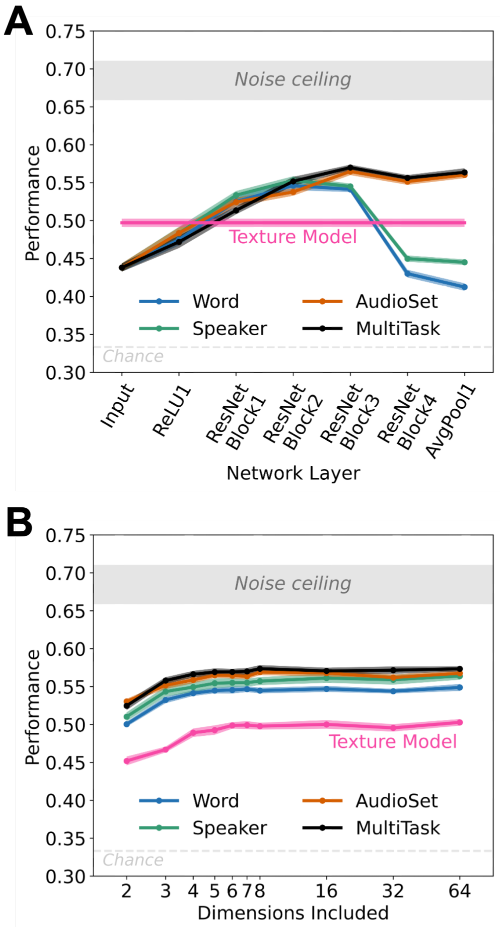

Figure 2: **A.** Model performance after optimizing a linear transformation of the representations from each model stage. **B.** Model performance for the best-performing stage after optimizing a linear projection to a low-dimensional feature space.

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. Finally, we found that 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.

Acknowledgments

Work supported by NSF grant BCS-2240406. The authors thank Martin Hebart for helpful discussions.

References

Feather, J., & McDermott, J. H. (2018, September). Auditory texture synthesis from task-optimized convolutional neural networks. In Conference on Computational Cognitive Neuroscience.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 776-780). IEEE.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nature human behaviour, 4(11), 1173-1185.

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. Current opinion in neurobiology, 55, 121-132.

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.

Roads, B. D., & Love, B. C. (2024). Modeling similarity and psychological space. Annual Review of Psychology, 75, 215-240.