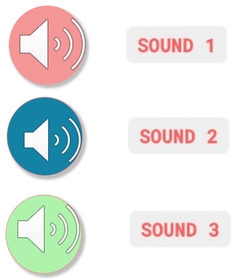
Discovering the Perceptual Space of Natural Sounds from Similarity Judgments

Ariana Richardson, Bryan Medina, Jarrod Hicks, and Josh McDermott

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

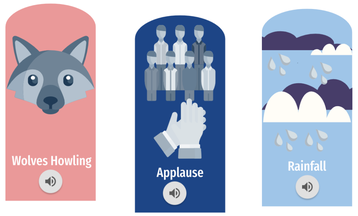
Introduction

- Understanding how the rich acoustic world gives rise to multidimensional mental representations is a challenge for perceptual science
- **How are human similarity judgments made between natural sounds?**

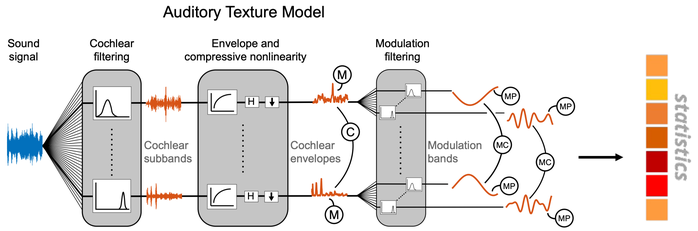


Sound Textures

- Sound textures are sounds created by a superposition of many similar acoustic events



- Thought to be represented perceptually in the form of time-averaged statistics

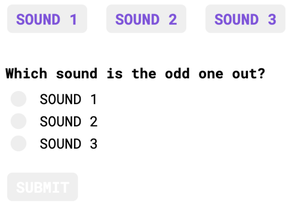


- However, it is unclear how statistics relate to perceptual similarity
- Model contains thousands of statistics

  →how does each statistic contribute to perception?

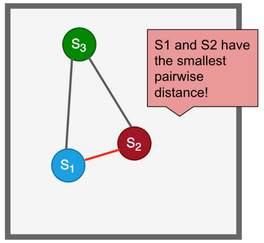
Data Collection

- Odd-one-out task to assess similarity
- Stimuli: 1080 2-second clips of sound textures
- Online data collection (N=117)



Observer Model

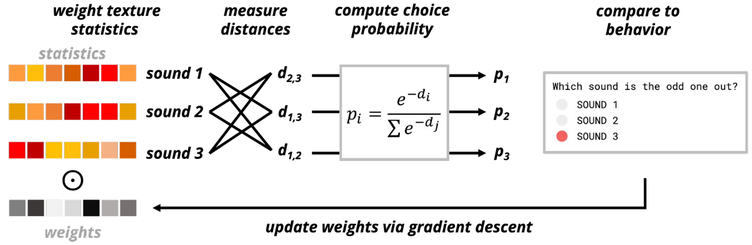
- Measures pairwise distances between representations of sounds, assuming a metric
- Sound not contained in minimum distance pair is chosen as the odd-one-out



- Model choices are compared to human choices in odd-one-out task

Model Fitting

- Weights were adjusted to maximize agreement with human participants



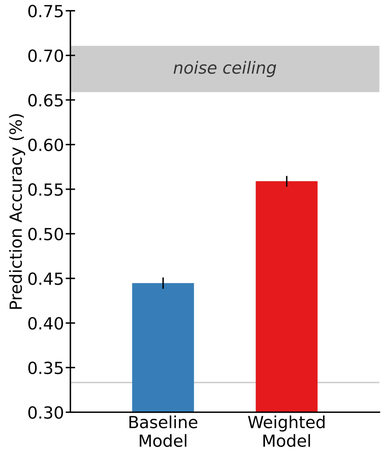
Results

- Model captures much of explainable variance in human similarity judgments

  →new perceptual metric for sounds
- But fitted model remains well below noise ceiling

  -could be missing acoustic features

  -or semantic associations that influence humans



Conclusions

- We developed a signal-computable computational model to predict human similarity judgments for real-world sound textures
- Model captures much of explainable variance in similarity judgments, achieving 64% of the best possible accuracy given the noise in the data.
- Current model of texture does not completely explain human similarity judgments.

Future Directions

- Add additional learnable dimensions to capture possible semantic features
- Use metric to model texture streaming and memory confusions