Contact: heesok@andrew.cmu.edu, jfeather@andrew.cmu.edu

Probing Invariances in Auditory Event Categorization Using Model Metamers

Hee So Kim[1], Elizabeth J. Lee[1], Malinda J. McPherson-McNato[2], Abigail L. Noyce[1], Jenelle Feather[1]

1 - Carnegie Mellon University, Pittsburgh, PA, USA. 2 - Purdue University, West Lafayette, IN, USA.

Carnegie Mellon University

Introduction

Invariance to acoustic features (e.g., pitch and timing) is essential for robust recognition of natural sounds

Deep neural networks (DNNs) achieve high performance on auditory event categorization tasks

→ Unclear if DNNs share human-like invariances

Questions:

(1) Do models and humans rely on similar invariant representations of auditory events?

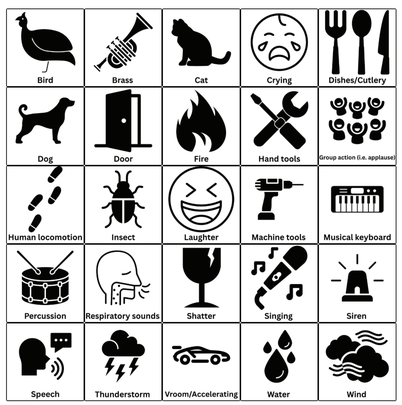
(2) How do invariances differ across model architectures and training objectives?

Approach:

Design a directly comparable task for humans and models

Evaluate human recognition of natural sounds and model metamers

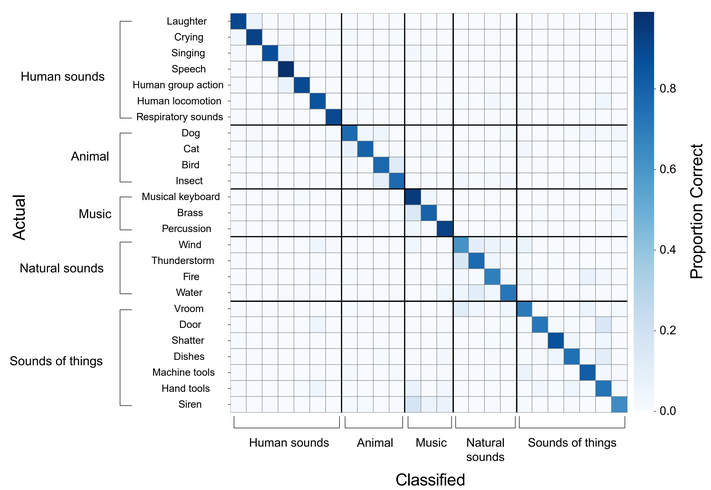
Humans correctly classify natural sounds in a novel classification paradigm



TASK:

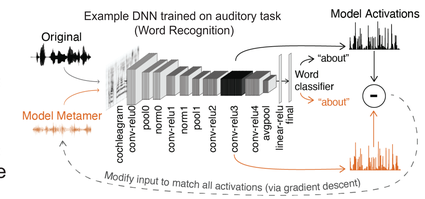
"Choose the category corresponding to the sound"

- Developed a novel sound classification task using 25 distinct categories, following structure of the AudioSet ontology [1]
- Selected 11 sound clips per category from the FSD50K dataset [2]
- All samples were 2s duration
- N=20 participants (Prolific)
- 80 +- 2% accuracy across all categories



Testing invariances with model metamers

- **Model metamers**: Two distinct stimuli that produce the same response in a computational model
- Human recognition of model metamers tests whether humans share model invariances
- Generated model metamers for a large set of high-performing publicly available audio models and a set of controlled in-house models with variations in training



Feather et al. 2023

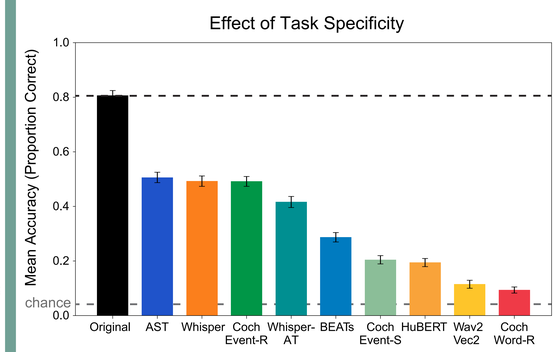
Models invariances are shaped by task specificity

Audio event recognition models:

- Audio Spectrogram Transformer (AST) [3]
- Bidirectional Encoder Audio Transformers (BEATs) [4]
- Whisper-AT [5]
- CochEvent-R
- CochEvent-S

Speech recognition models:

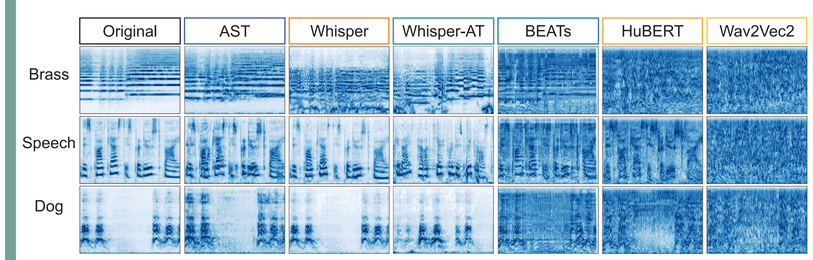
- HuBERT [6]
- Wav2Vec2.0 [7]
- Whisper [8]
- CochWord-R [9]



Tested model metamers generated from the final representation layer of each model

Human-like invariances for sound categorization emerge in models trained on auditory events

- Models trained on speech recognition generally have non human-like invariances for auditory event categorization
- Exception: Humans are able to recognize metamers from Whisper

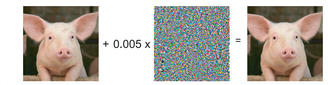


Conclusion

- Developed a paradigm testing model-human invariance alignment in audio event categorization
- Metamer recognizability increases with task specificity and adversarial training
- At the classification stage, models still have non human-like invariances

Human recognition of model metamers increases with adversarial training

- Adversarial training improves alignment to human perception by increasing robustness to small cochleagram perturbations [9, 10]

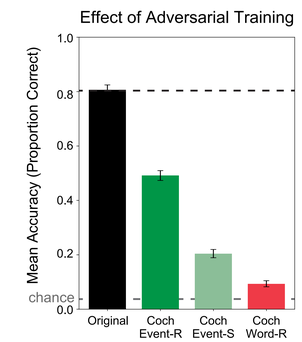


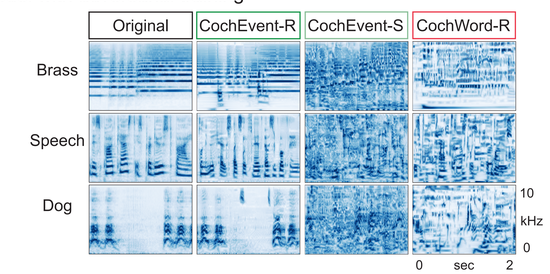
"pig" → "plane"

Goodfellow et al. 2015

In-house models (Cochleagram + ResNet50):

- **CochEvent-R:** Auditory event recognition model with adversarial training
- **CochEvent-S:** Auditory event recognition model with standard training
- **CochWord-R:** Speech recognition model with adversarial training

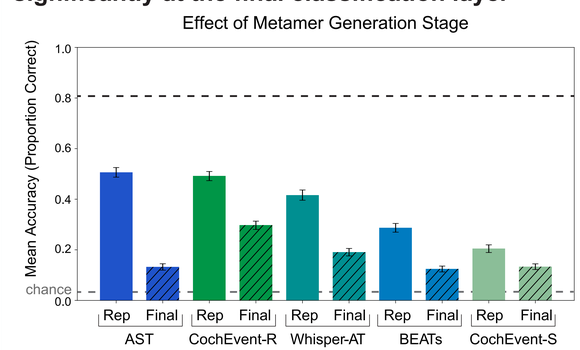




Tested model metamers generated from the final representation layer for a set of controlled in-house models

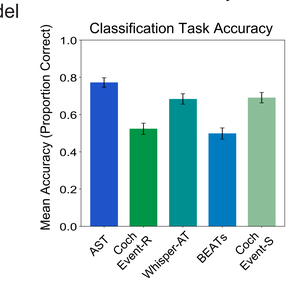
- Adversarial training improves recognizability of model metamers, but only when model is trained on an auditory event task

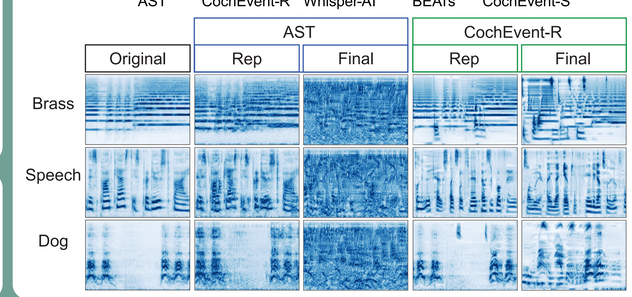
Recognizability of model metamers for audio event reognition models drop significantly at the final classification layer



Tested model metamers generated from the final representation layer and the final classification layer for models trained on audio event recognition

- Recognizabiltiy of model metamers drops significantly at the final classification stage
- Decrease in recognizability is least pronounced in the adversarially trained model





Model Metamer Sound Examples

References: [1] Gemmeke et al. 2017 [2] Fonseca et al. 2021 [3] Gong et al. 2021 [4] Chen et al. 2022 [5] Gong et al. 2023 [6] Hsu et al. 2021 [7] Baevski et al. 2020 [8] Radford, Kim et al. 2023 [9] Feather et al. 2023 [10] Goodfellow et al. 2015

Funding: This work is supported in part by MURI N00014-19-12332