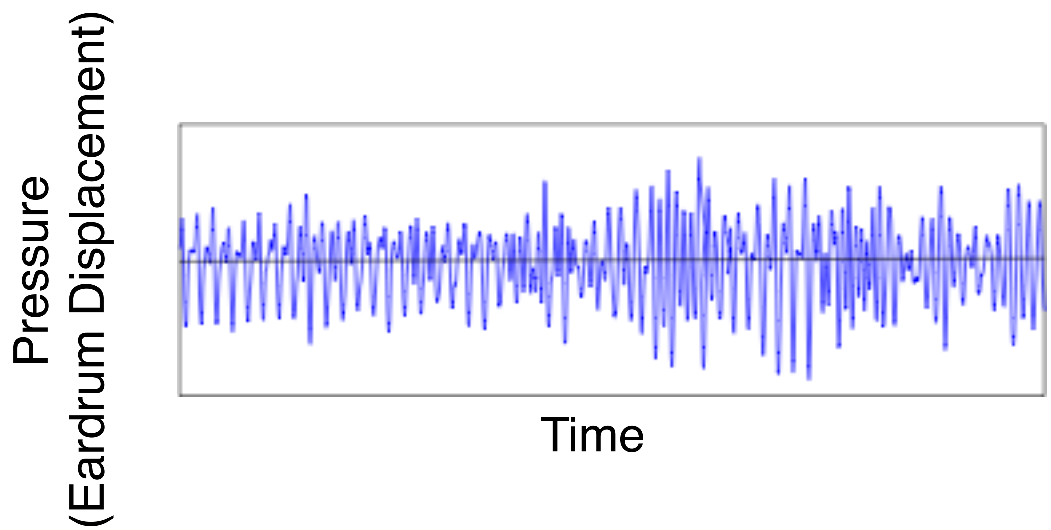
New Models of Human Hearing via Machine Learning

Josh McDermott
Dept. of Brain and Cognitive Sciences, MIT
McGovern Institute for Brain Research, MIT
Center for Brains, Minds, and Machines, MIT
Program in Speech and Hearing Biosciences and Technology, Harvard

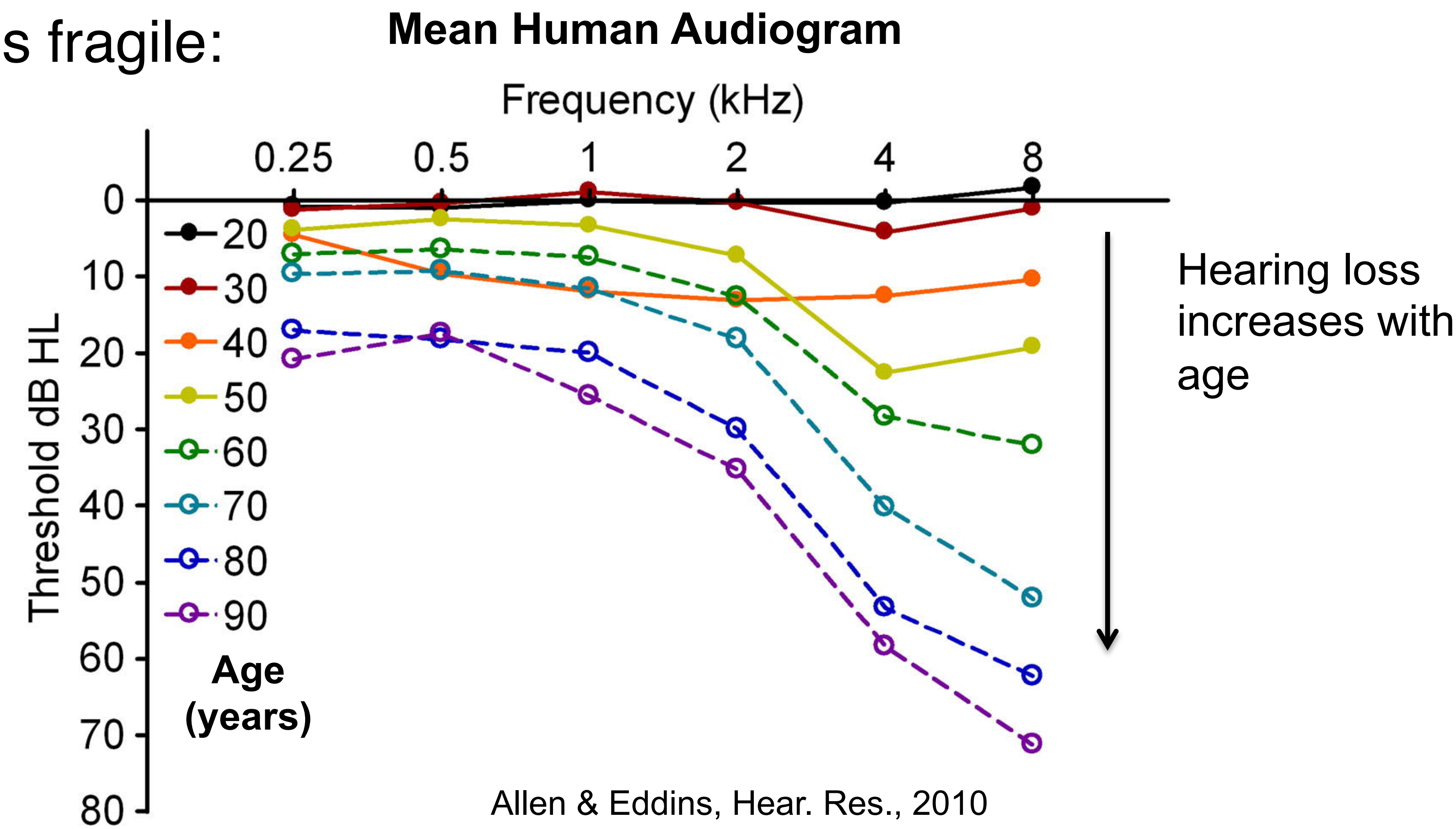
Everyday human listening is a stunning computational feat…

Consider an example of typical auditory input:

The ear receives a pressure waveform.

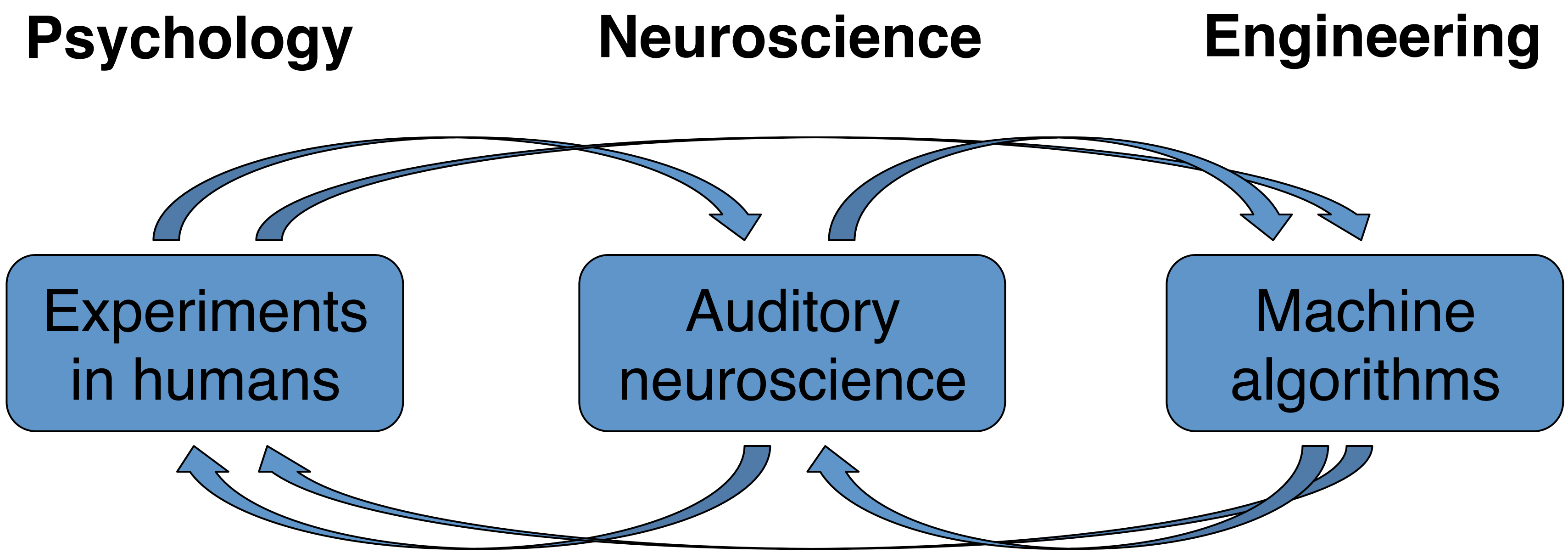


Hearing is fragile:



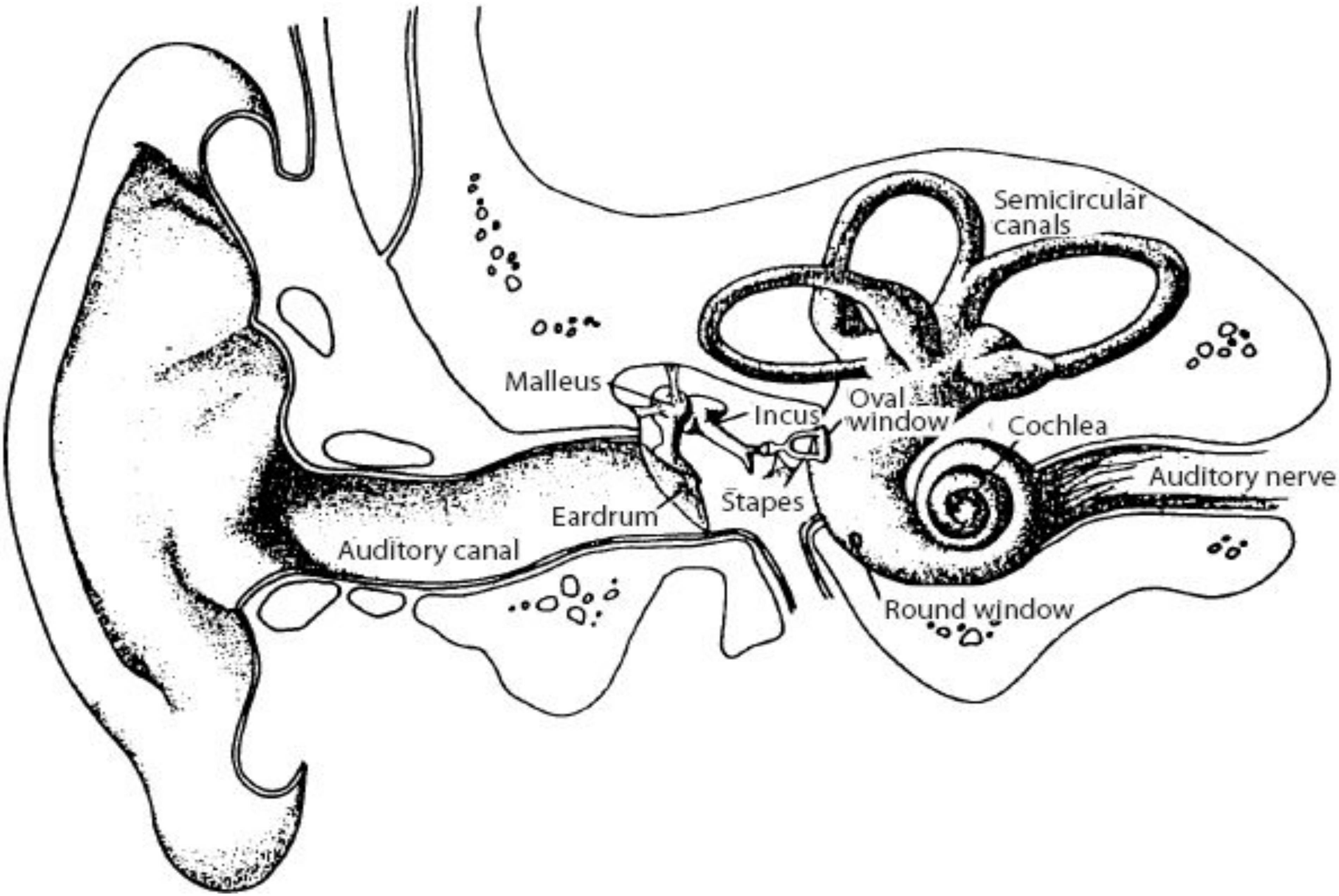
- Current hearing aids help in quiet, less so in noisy environments
- Limited by our understanding of how we hear

Our research group: Laboratory for Computational Audition

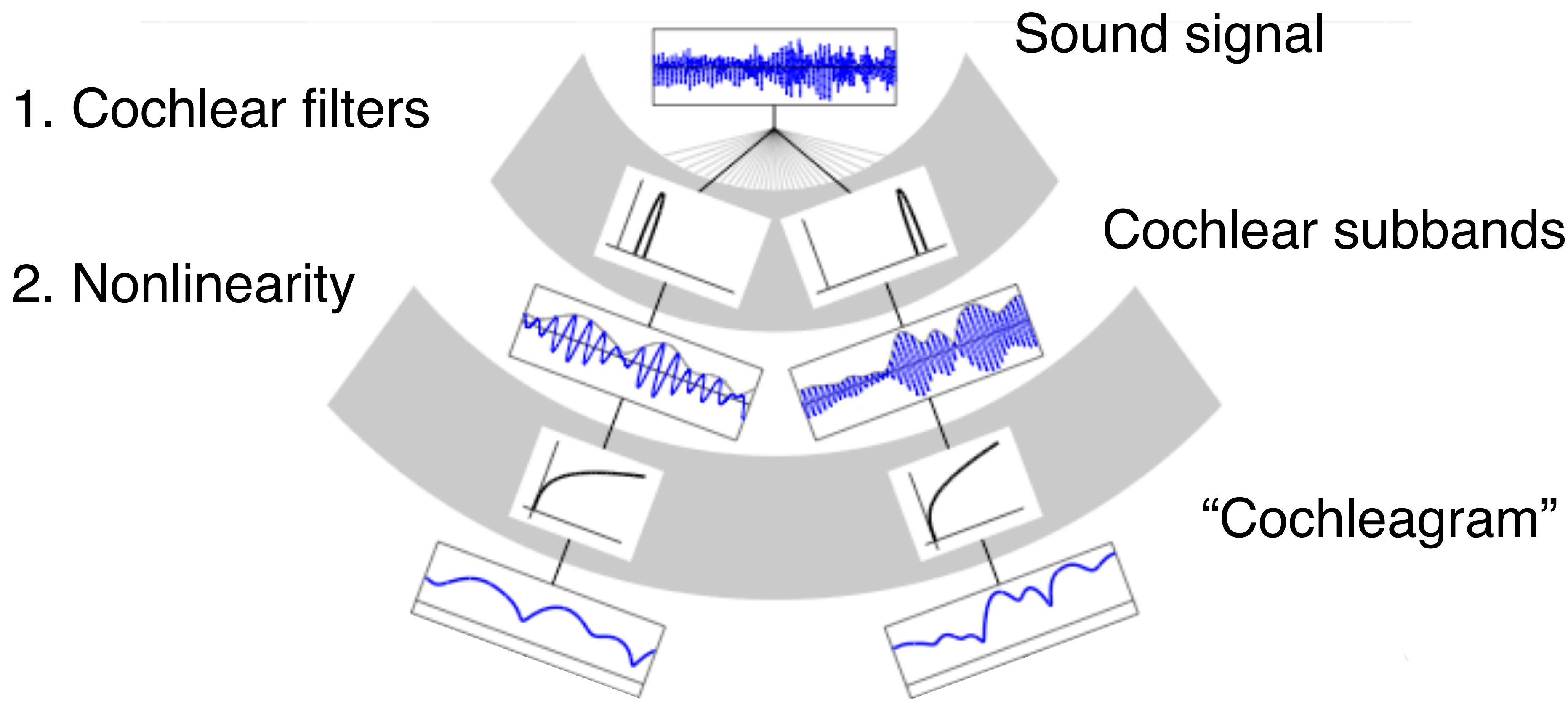


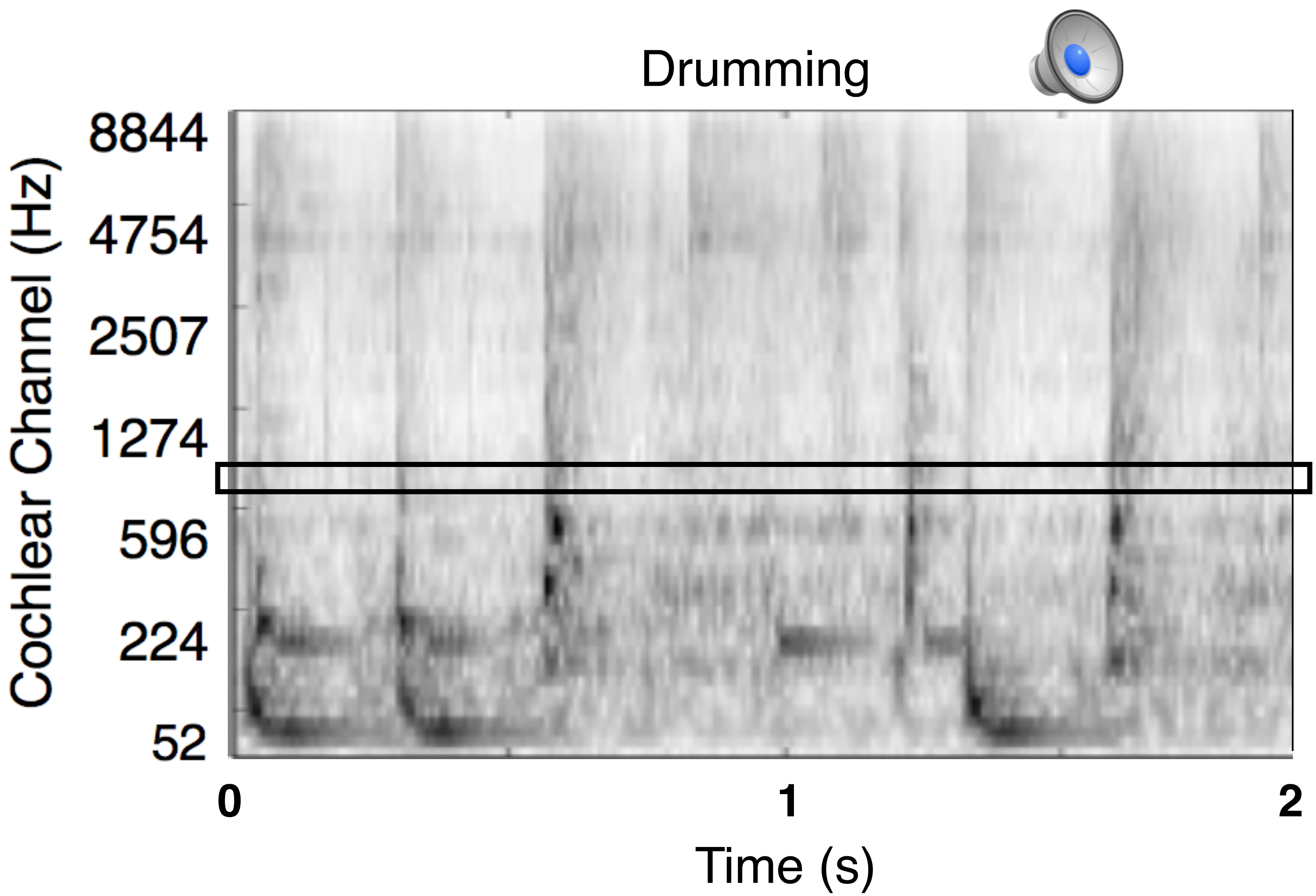
- Goal: to build good predictive models of human hearing
- If successful, will transform our ability to make people hear better

Peripheral auditory system is fairly well characterized.



Standard peripheral auditory models:



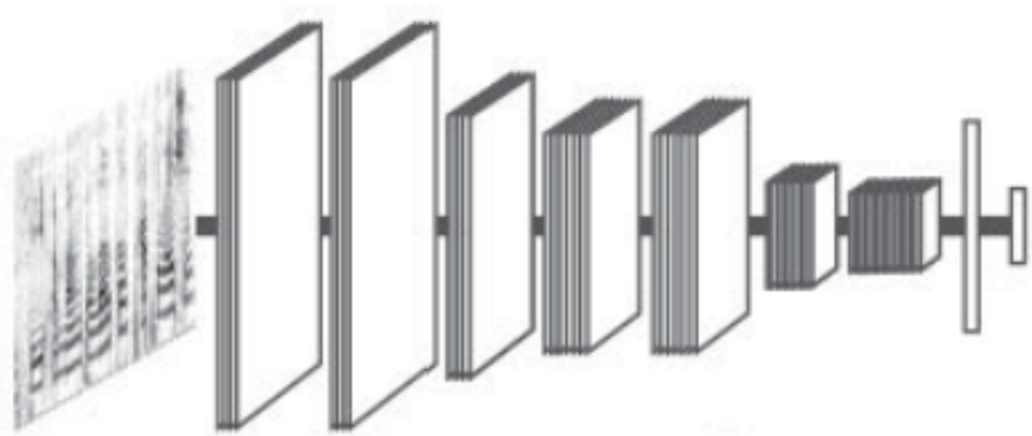

Drumming
Cochlear Channel (Hz)
8844
4754
2507
1274
596
224
52
0
1
2
Time (s)

What happens downstream?

Can we obtain better models by training systems to perform tasks?

Can we obtain better models by training systems to perform tasks?

Human-level performance on classification tasks is now routine via artificial neural networks



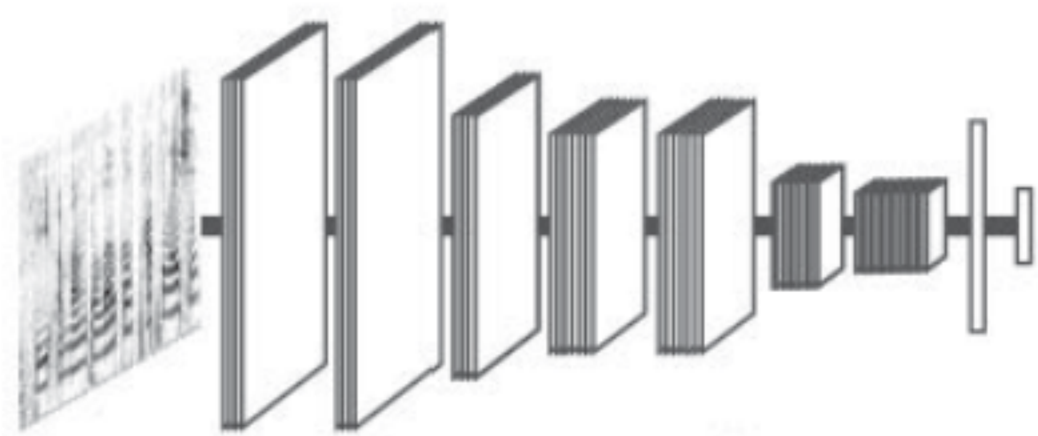
Repeated application of simple operations:

filtering (convolution), pooling, and normalization

Filters and model architecture can be optimized to classify input signal

Can we obtain better models by training systems to perform tasks?

- Hardwire cochlea to be faithful to biology
- Learn all subsequent stages with a neural network



Result: Candidate model of auditory system

Many widely discussed limitations:

- Learning is unrealistic…
- "Neural" networks are not very neural…
  - Poorly suited to circuit-level models
- Behavior typically limited to trained classification tasks

But for now:

Deep learning enables optimization of hierarchical models for real-world tasks.

→ optimized observer models in new domains.

Plan for Today

- Summary of recent successes of our neural network models of hearing
- Discussion of current model shortcomings

Take-Home Messages, Part 1

After training on natural auditory tasks with natural sounds:

- Pretty good matches to human behavioral experiments
    - Speech recognition in noise
    - Sound localization
    - Pitch perception

- Best current predictions of auditory cortical responses

Manipulation of training conditions shows that similarity is a function of optimization for natural tasks/sounds, cochlea

- Provides insight into origins of human behavioral traits

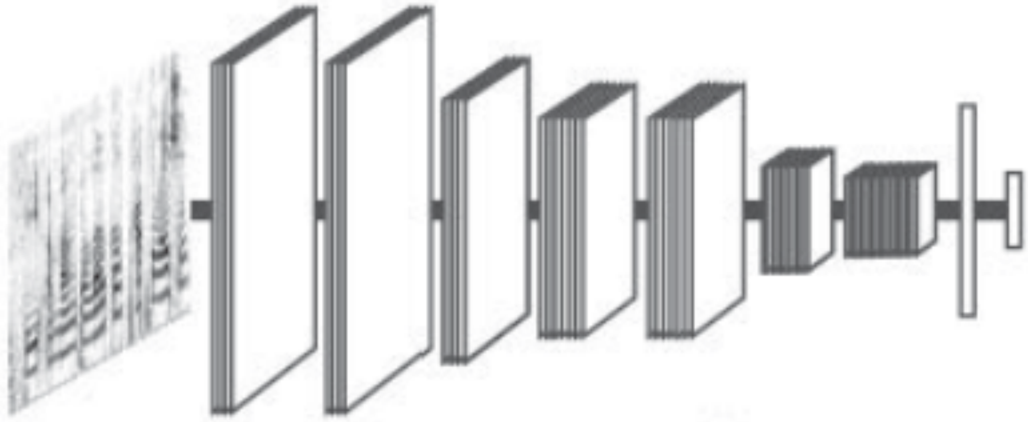
Degrading simulated cochlear input to the neural network reproduces characteristics of human hearing impairment

SPEECH RECOGNITION IN BACKGROUND NOISE

Excerpted speech
+
Background noise
(e.g., music, speech babble, auditory scenes)

"…gross domestic product grew…"
2 sec.

What word occurred halfway through clip?
600-way classification task



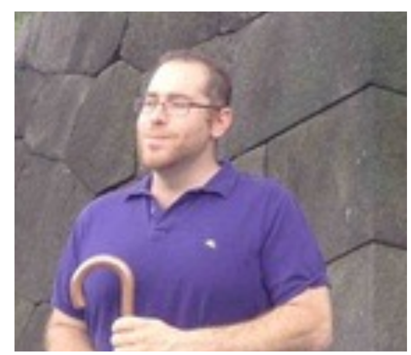
Classify word

- Weights learned with standard backpropagation

- Automated optimization of architectural hyperparameters

- Convolutional in time and frequency

- Sounds are relatively short (< 2s), so we neglect directionality of time, memory etc.



Kell et al., Neuron, 2018

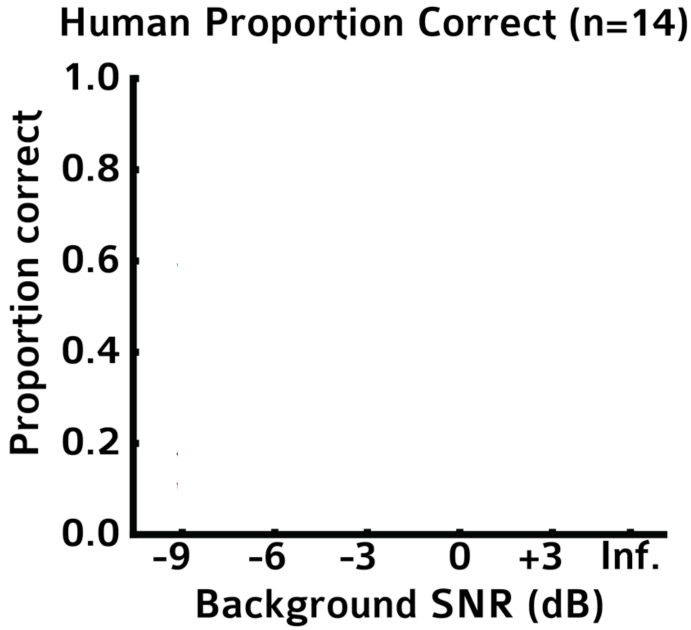
Behavioral comparison:
Speech recognition in background noise

<u>21 conditions:</u>
clean
+
4 different background types at 5 SNR levels

600 AFC

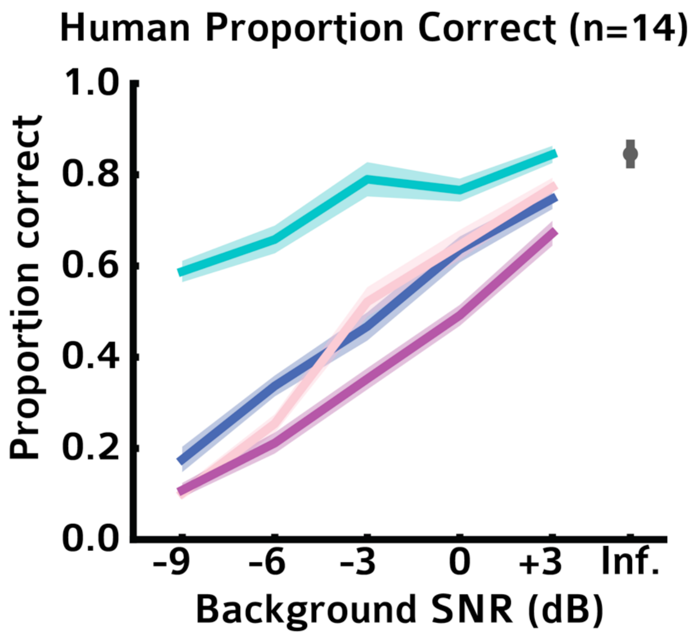
Erica Shook

Behavioral comparison: CNN & humans on same task



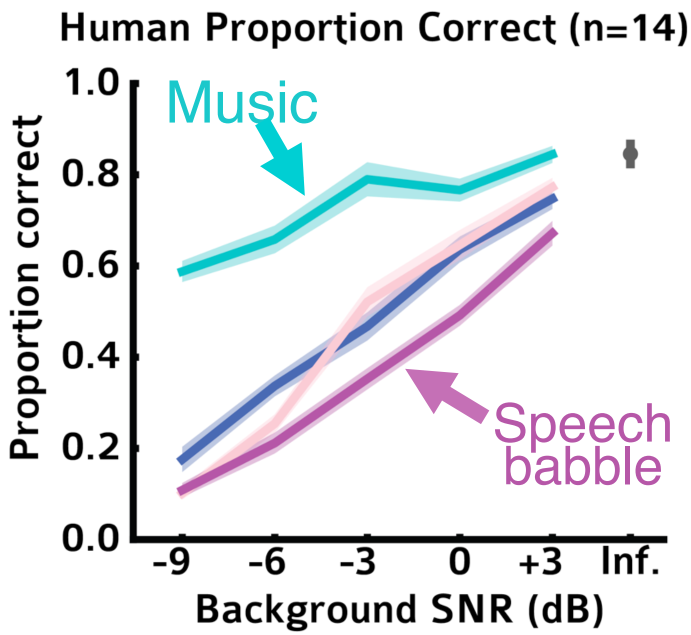
Music
Auditory scenes
Speech babble
Speech-shaped noise

Behavioral comparison: CNN & humans on same task



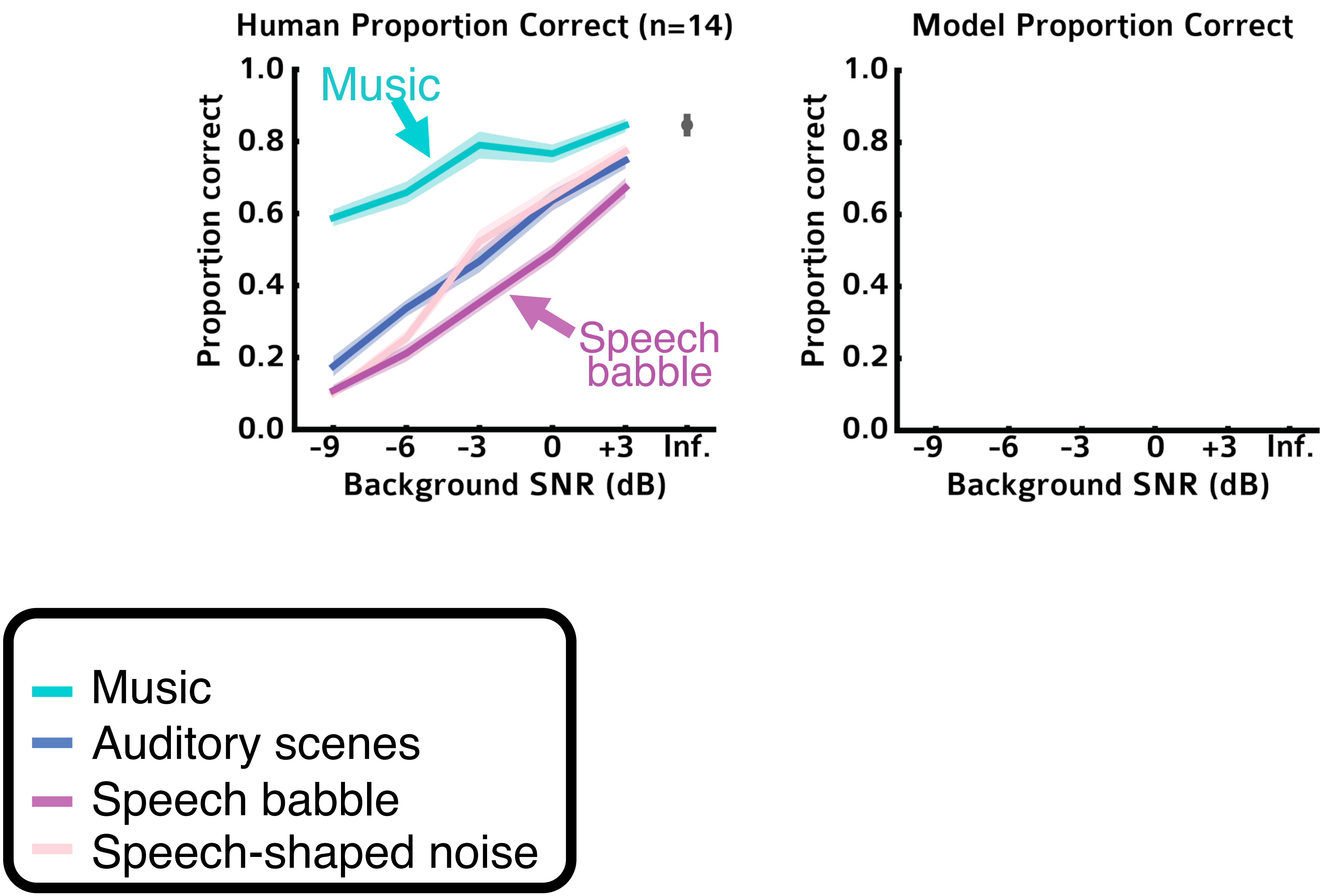
Music
Auditory scenes
Speech babble
Speech-shaped noise

Behavioral comparison: CNN & humans on same task

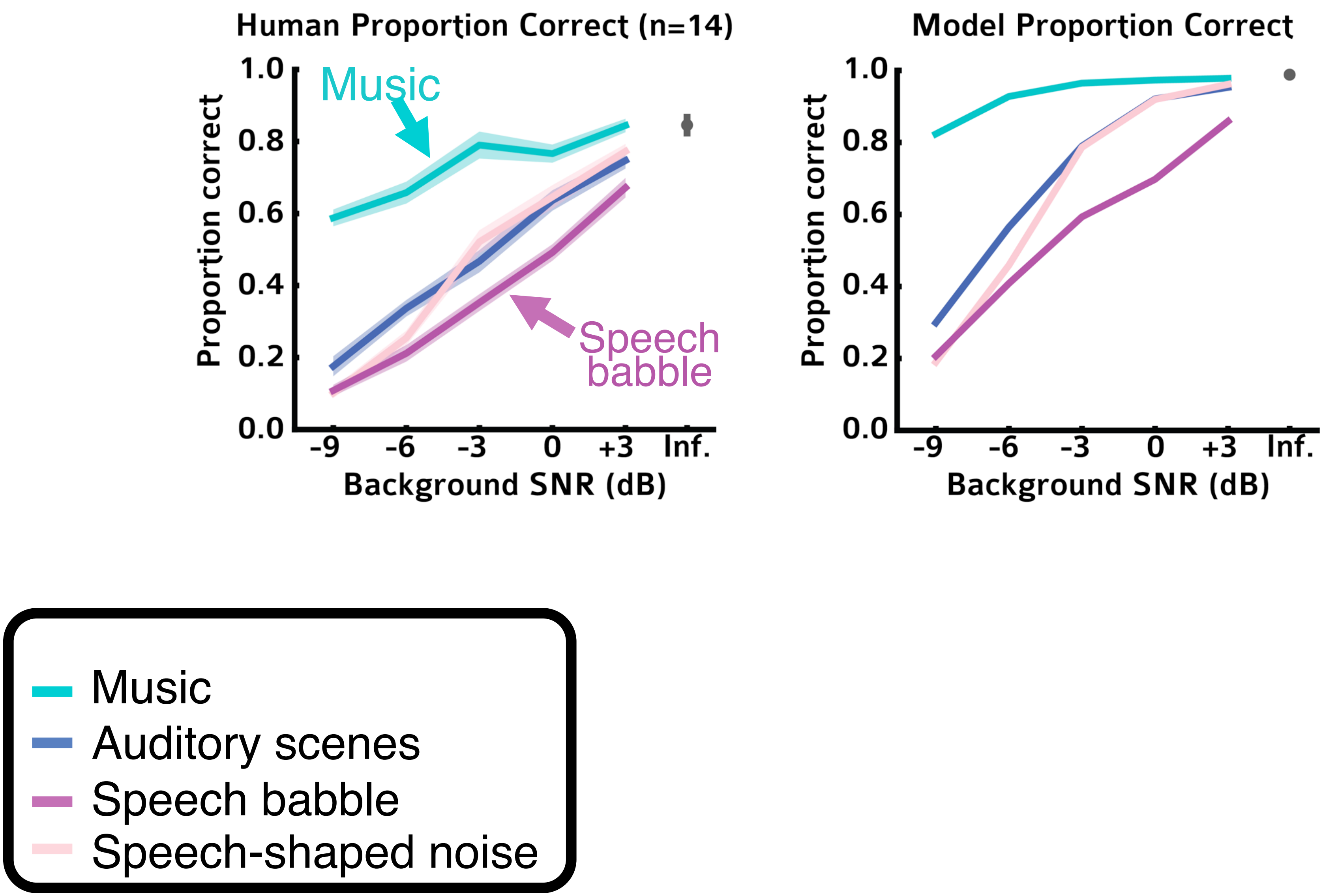


Music
Auditory scenes
Speech babble
Speech-shaped noise

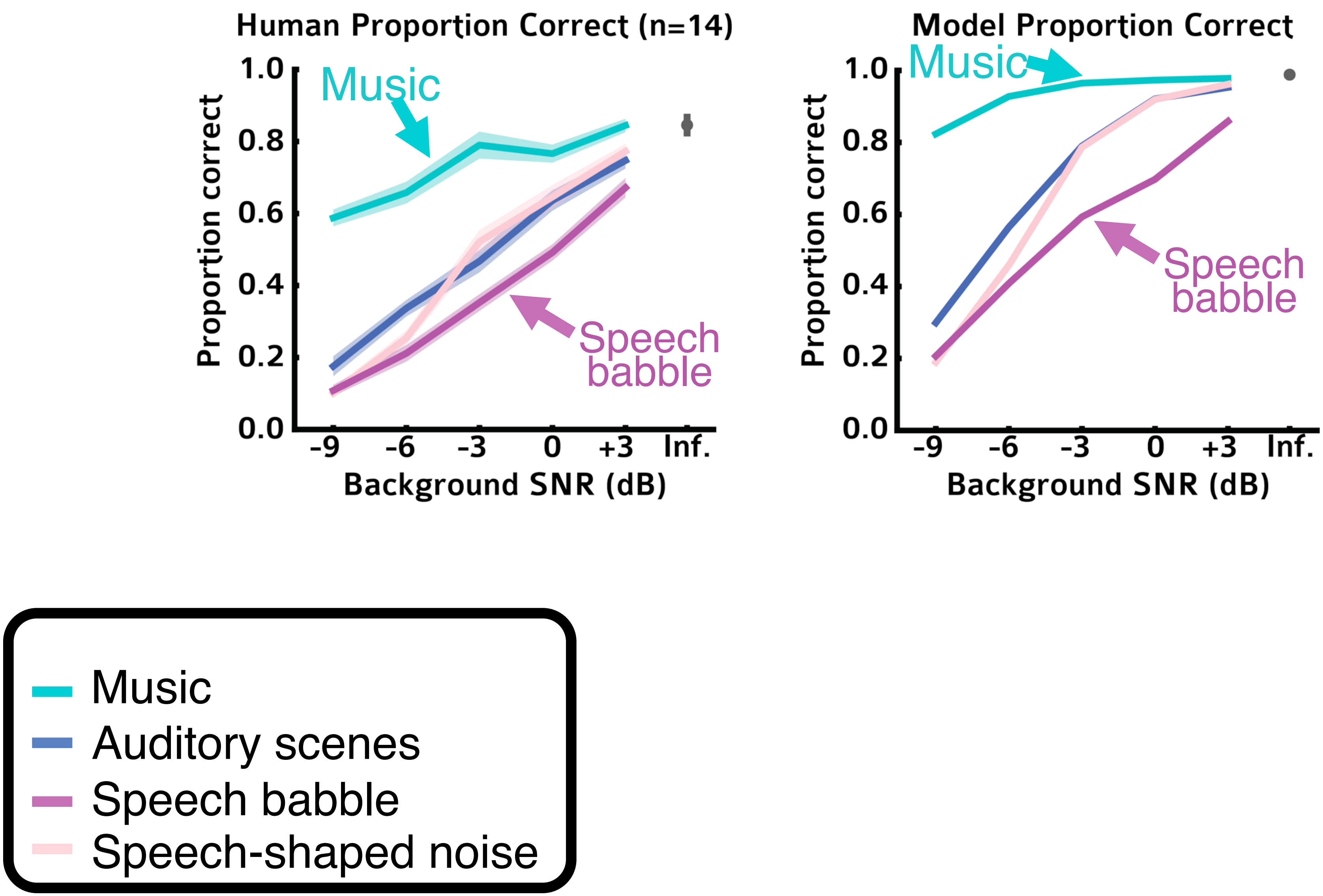
Behavioral comparison: CNN & humans on same task



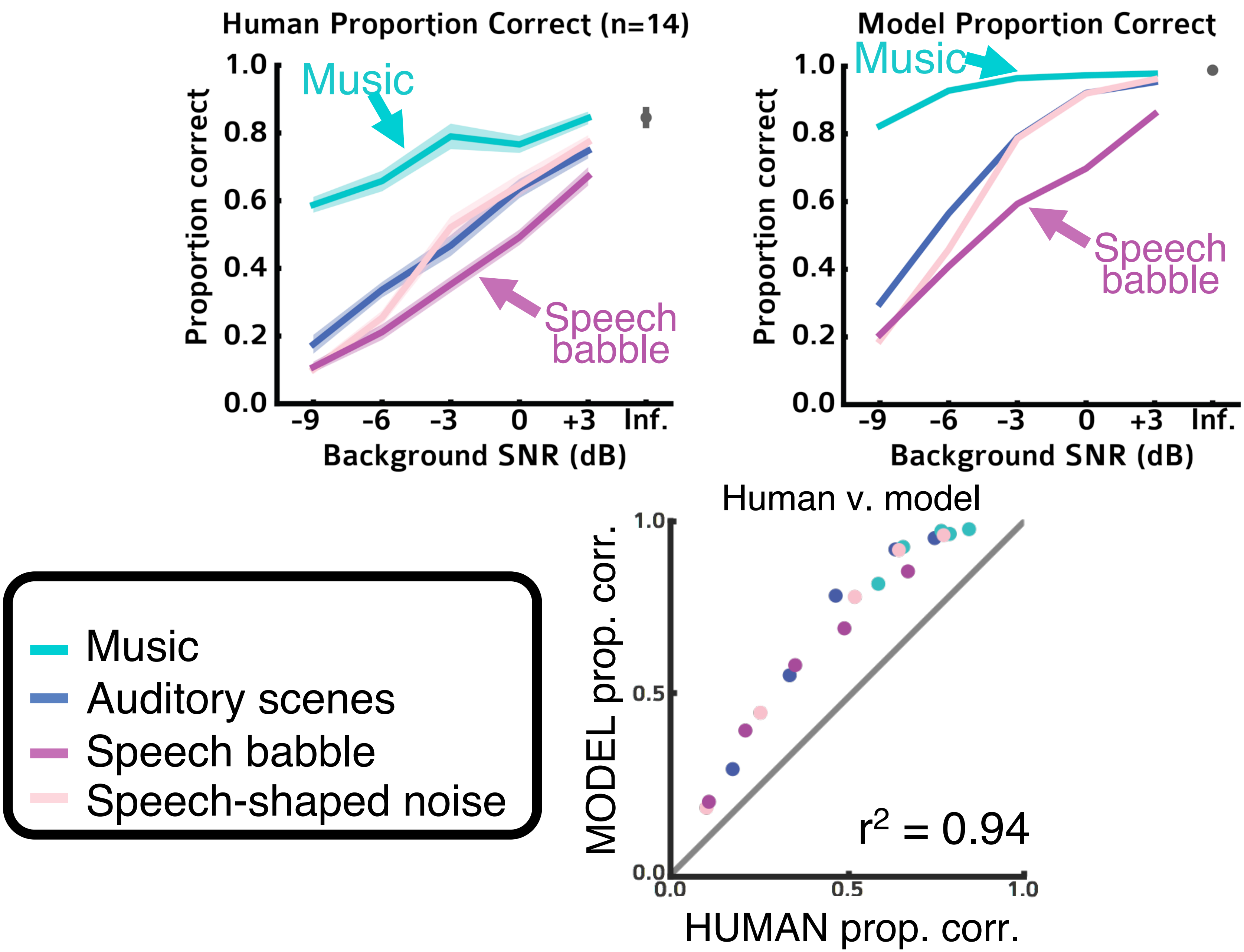
Behavioral comparison: CNN & humans on same task



Behavioral comparison: CNN & humans on same task

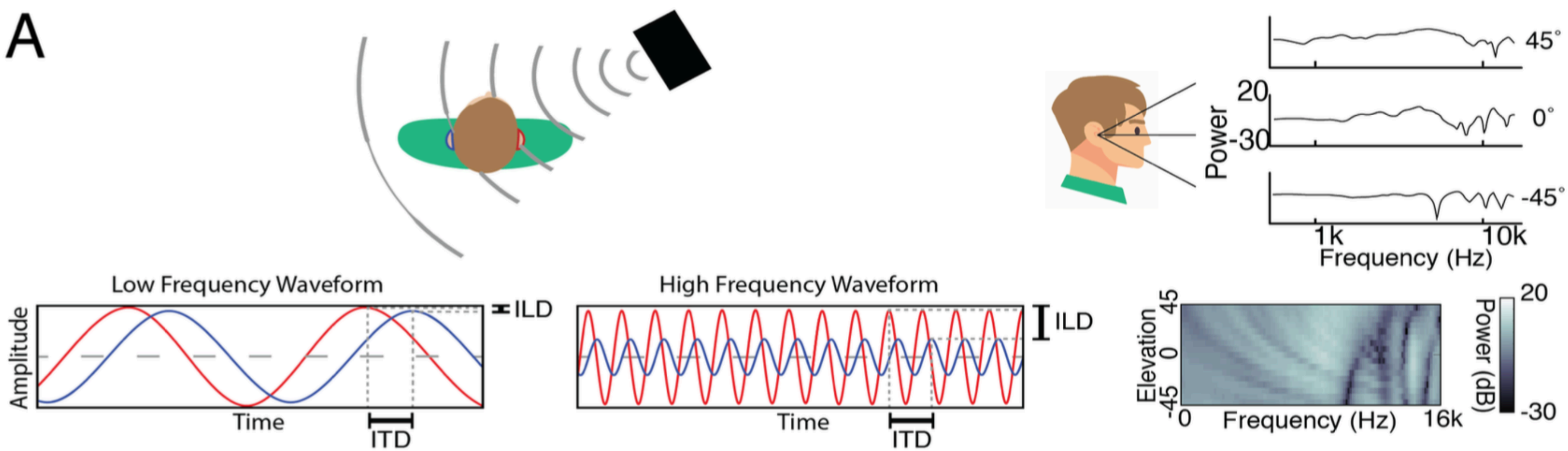


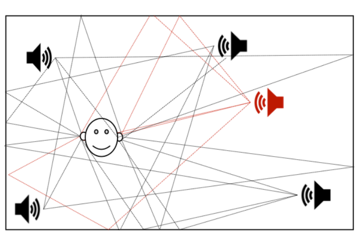
Behavioral comparison: CNN & humans on same task



Behavioral comparison: Sound localization

Classical story: three main types of cues to a sound's location



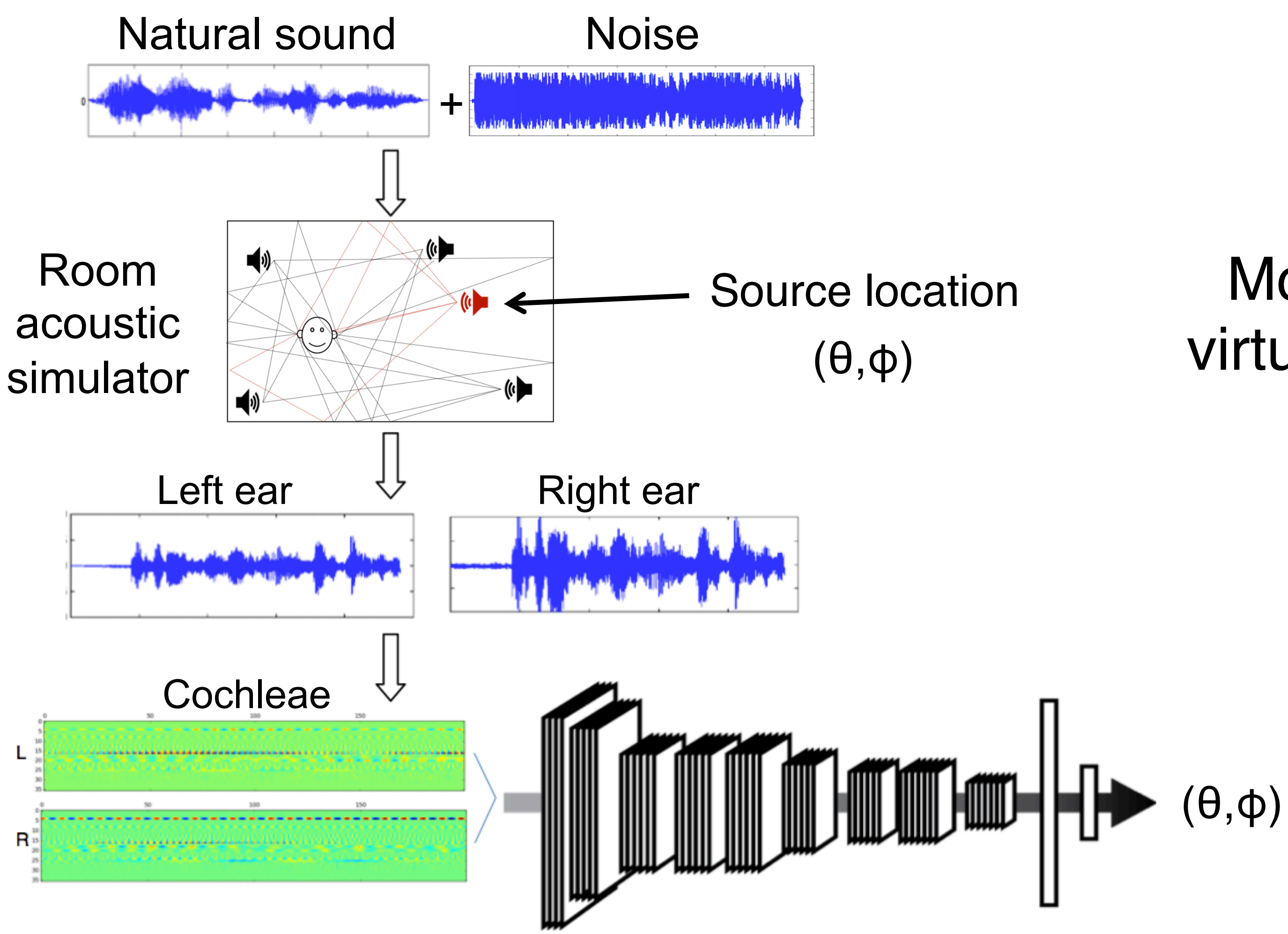


But real-world environments have noise, and reflections…

→ Hard problem

→ Models usually can't actually localize sounds

Behavioral comparison: Sound localization

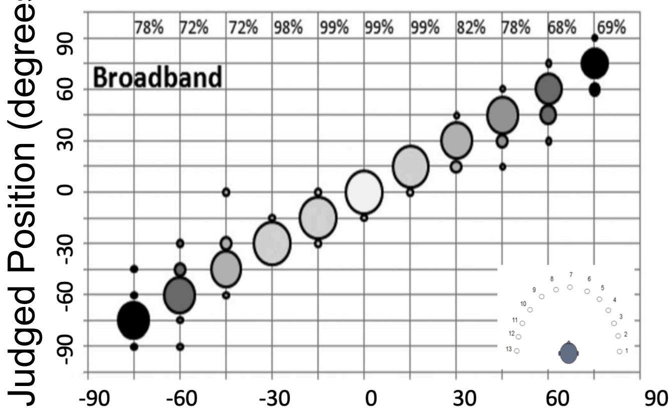


Model trained in virtual environment

Andrew Francl

Behavioral comparison: Sound localization

Human Localization

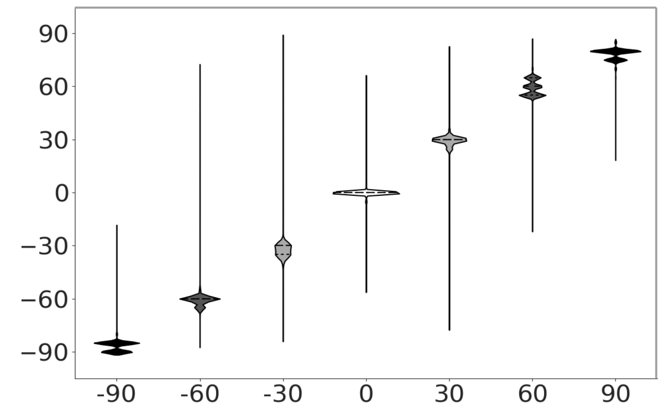


Actual Position (degrees)

Yost et al., JASA, 2013

Recordings from mannequin ears

Model Localization

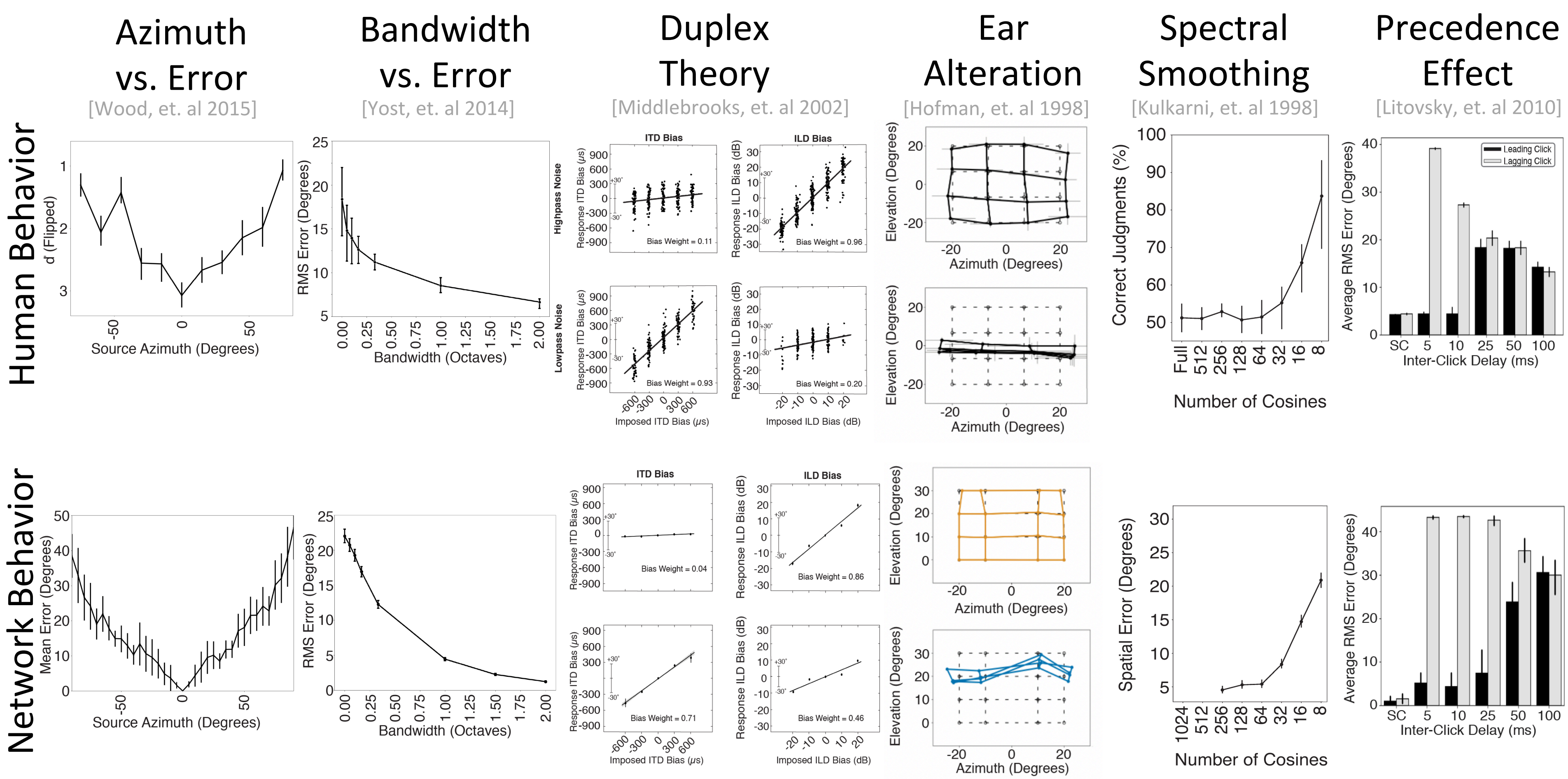


Actual Position (degrees)

Generalizes to
real-world (our
lab space at MIT)

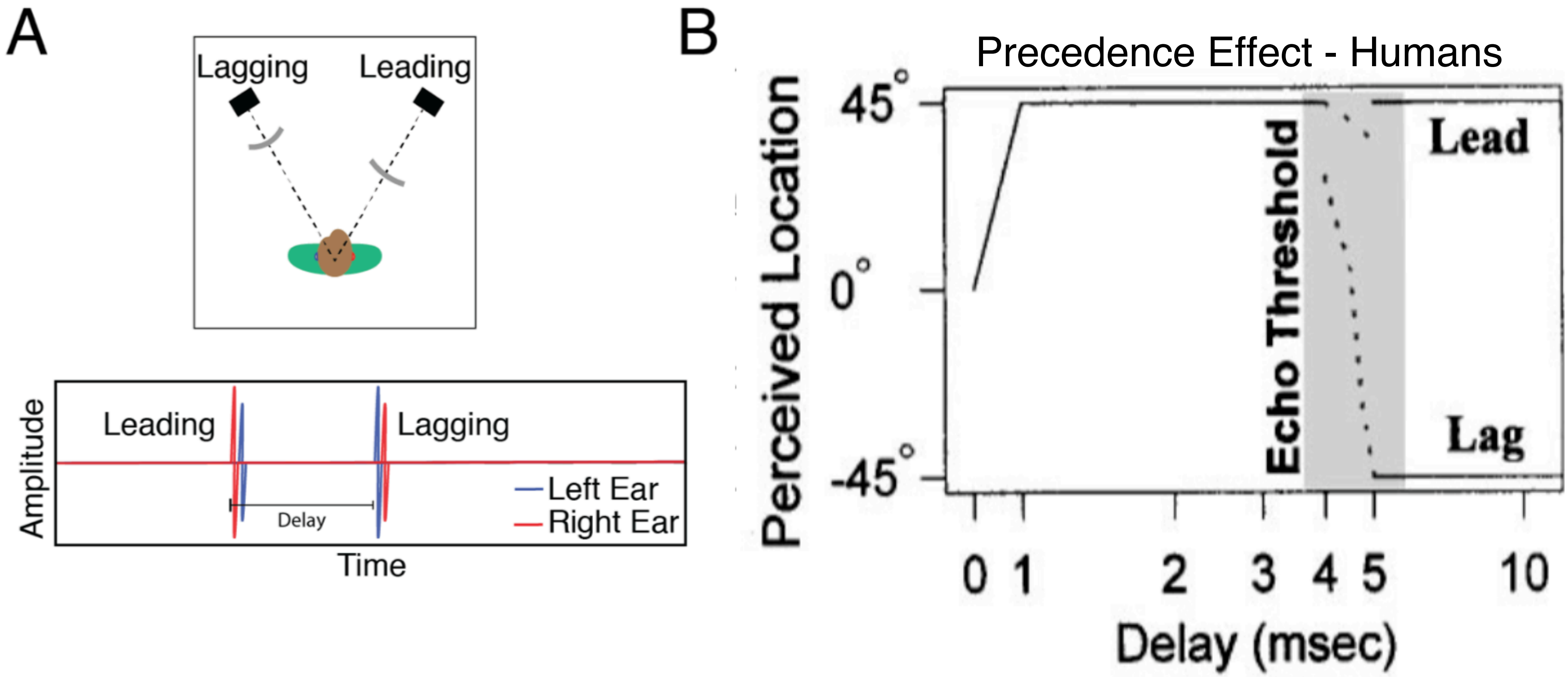
Andrew Francl

Trained network reproduces many properties of spatial hearing:



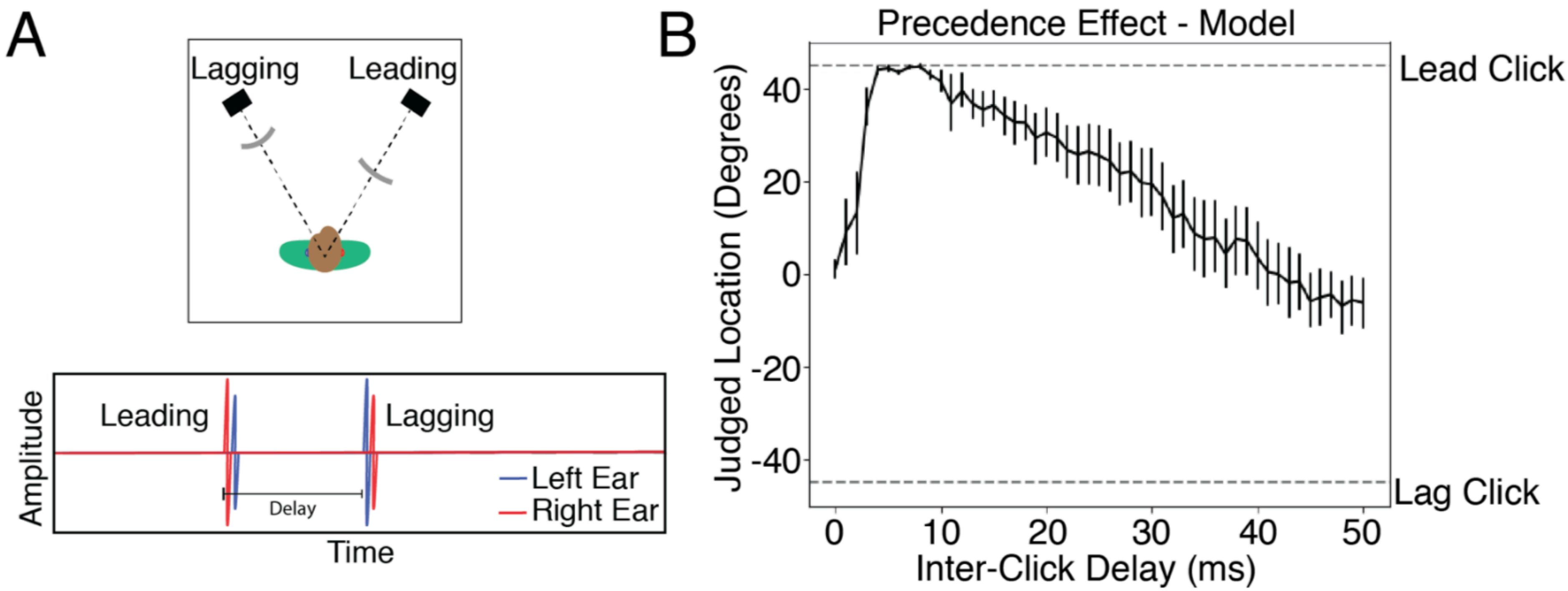
https://www.biorxiv.org/content/10.1101/2020.07.21.214486v1

Network's judgments are dominated by sound onsets ('precedence effect'), like humans:



https://www.biorxiv.org/content/10.1101/2020.07.21.214486v1

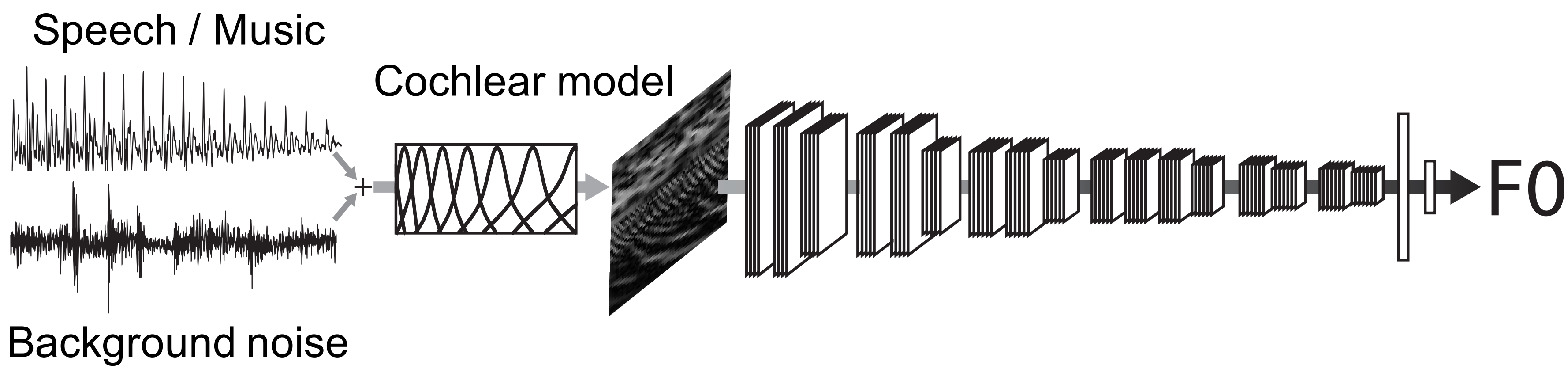
Network's judgments are dominated by sound onsets ('precedence effect'), like humans:



Check out the pre-print for lots of other examples:

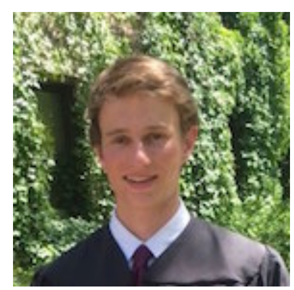
https://www.biorxiv.org/content/10.1101/2020.07.21.214486v1

Behavioral comparison: Pitch perception

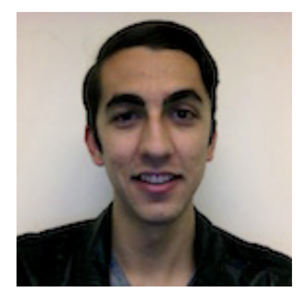


What was the fundamental frequency of the sound?

Network trained on speech, instruments in noise



Mark Saddler

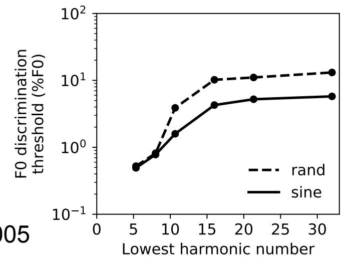


Ray Gonzalez

Effect of

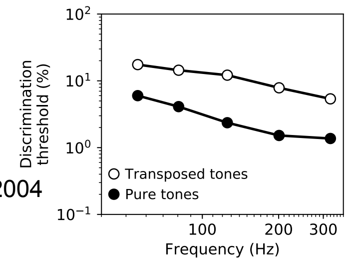
Human

harmonic number and phase



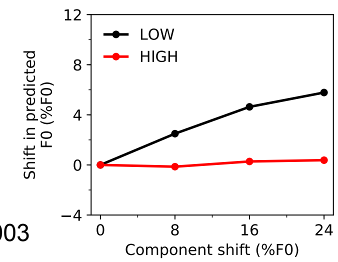
Bernstein et al., JASA, 2005

transposed tones



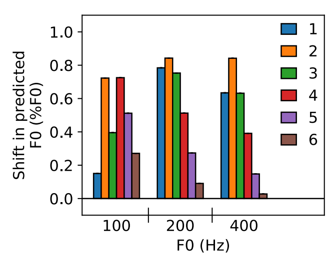
Oxenham et al., PNAS, 2004

frequency shifting of harmonics



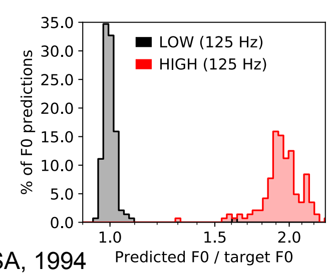
Moore & Moore, JASA, 2003

mistuning individual harmonics



Moore et al., JASA, 1985

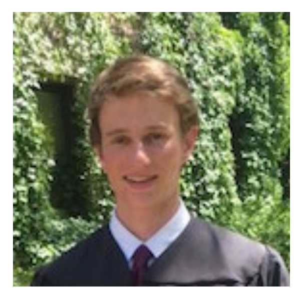
alternating harmonic phase

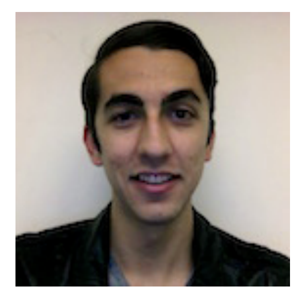


Shackleton & Carlyon, JASA, 1994

Assortment of classic behavioral characteristics of pitch (synthetic stimuli not in training set)

Does network replicate classic psychoacoustic pitch results?





Mark Saddler Ray Gonzalez

Effect of

harmonic number and phase

Bernstein et al., JASA, 2005

transposed tones

Oxenham et al., PNAS, 2004

frequency shifting of harmonics

Moore & Moore, JASA, 2003

mistuning individual harmonics

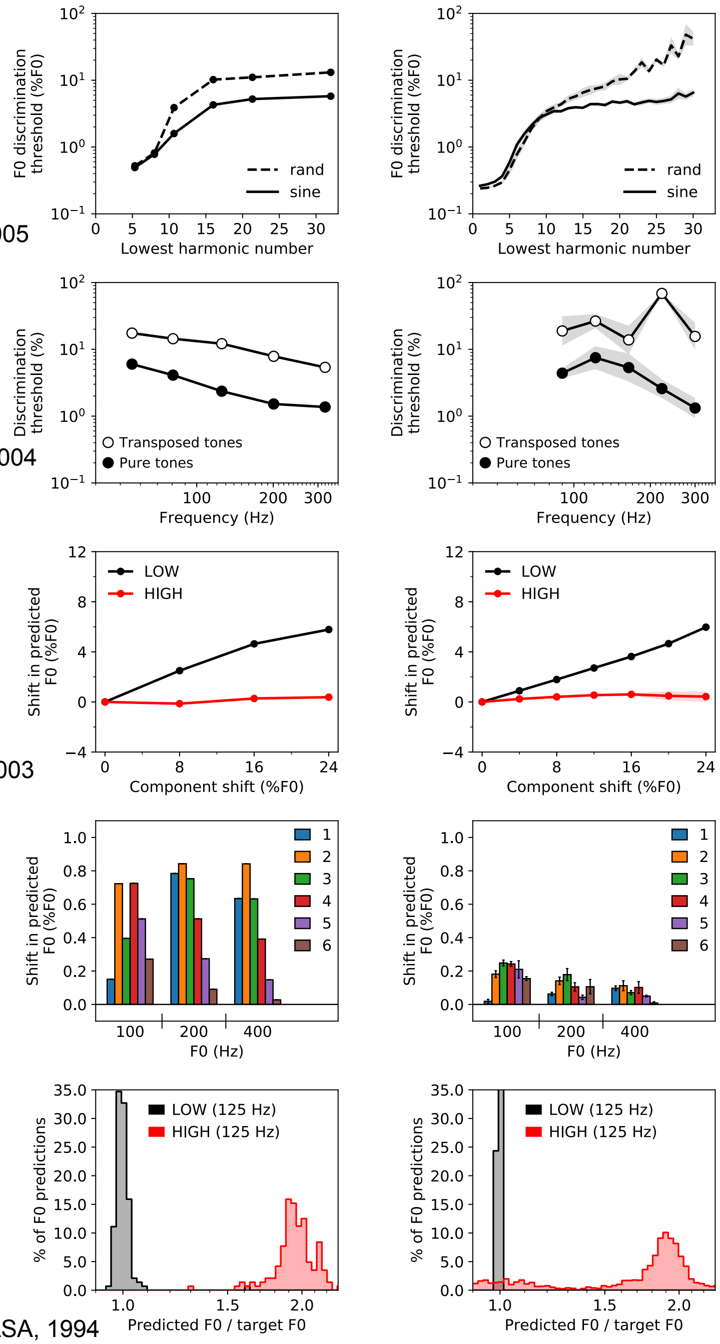
Moore et al., JASA, 1985

alternating harmonic phase

Shackleton & Carlyon, JASA, 1994

Human

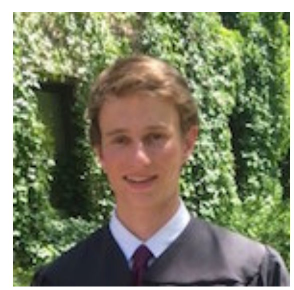
Model



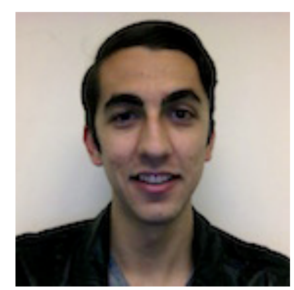
Assortment of classic behavioral characteristics of pitch (synthetic stimuli not in training set)

Does network replicate classic psychoacoustic pitch results?

Network reproduces key properties of human pitch perception.



Mark Saddler

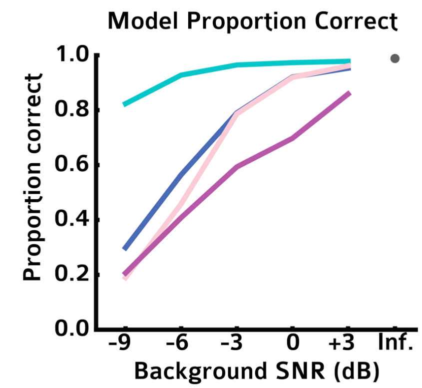


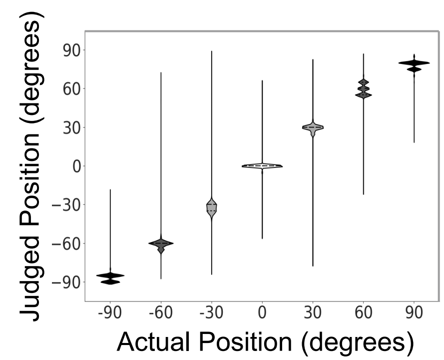
Ray Gonzalez

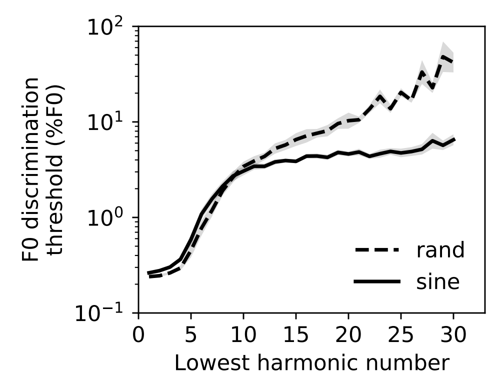
Major advance over previous models: human-like behavior

- In realistic conditions
- Comparable accuracy
- Similar psychophysics
  - → Similar use of cues

Allows investigation of conditions that produce human-like behavior







Effect of

harmonic number and phase

Bernstein et al., JASA, 2005

transposed tones

Oxenham et al., PNAS, 2004

frequency shifting of harmonics

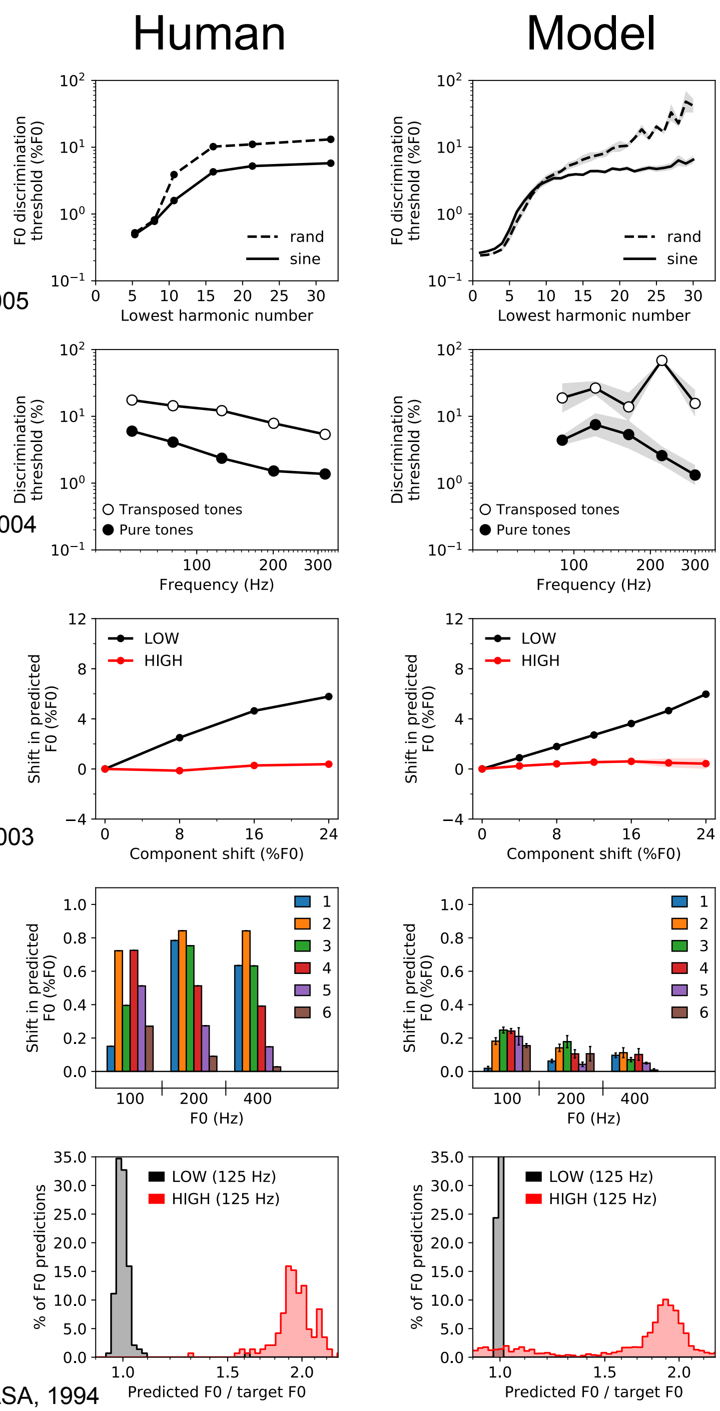
Moore & Moore, JASA, 2003

mistuning individual harmonics

Moore et al., JASA, 1985

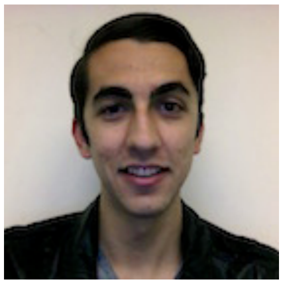
alternating harmonic phase

Shackleton & Carlyon, JASA, 1994



Model trained on natural sounds reproduces human characteristics.

To test whether learned strategy is adapted to natural environment, we instead train on unnatural synthetic tones (here with highpass spectra).



Mark Saddler Ray Gonzalez

Effect of

Human

Model trained on unnatural sounds

harmonic number and phase

Bernstein et al., JASA, 2005

transposed tones

Oxenham et al., PNAS, 2004

frequency shifting of harmonics

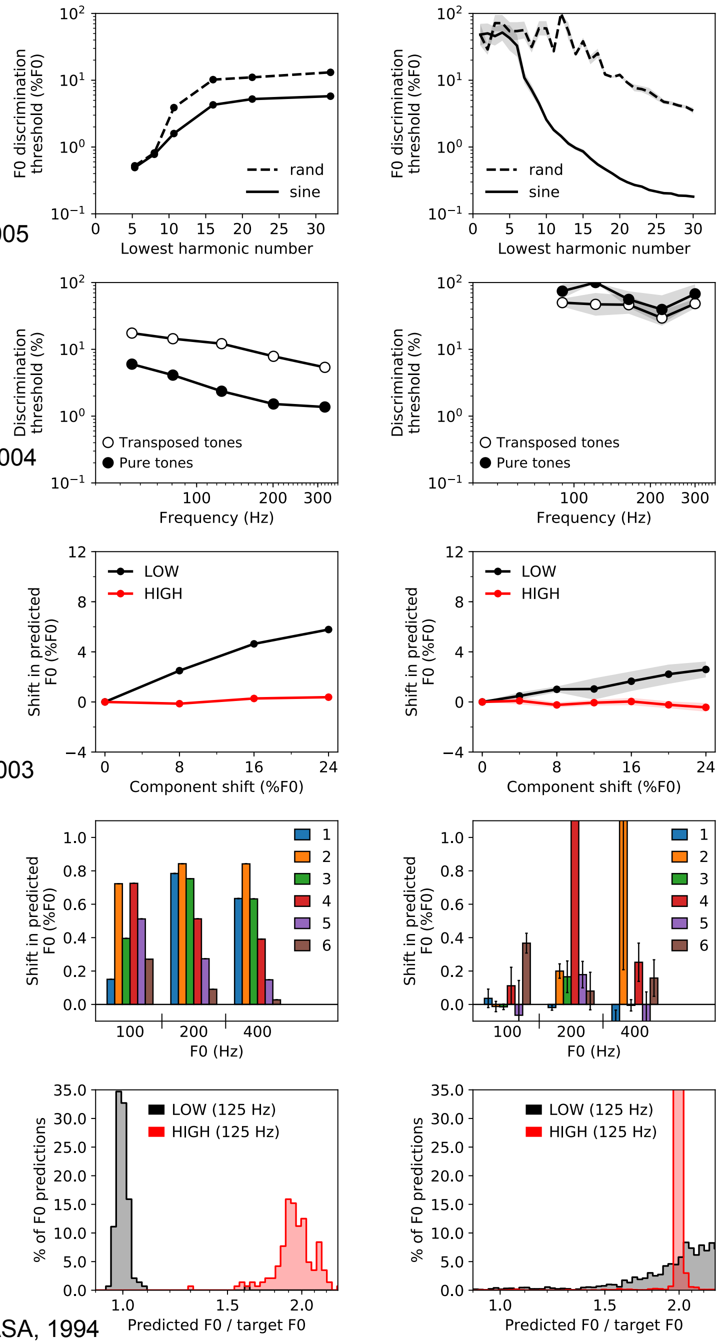
Moore & Moore, JASA, 2003

mistuning individual harmonics

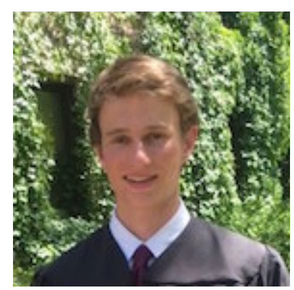
Moore et al., JASA, 1985

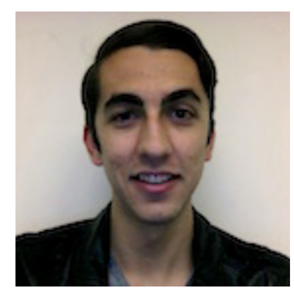
alternating harmonic phase

Shackleton & Carlyon, JASA, 1994



Model only resembles humans if optimized for natural sounds.

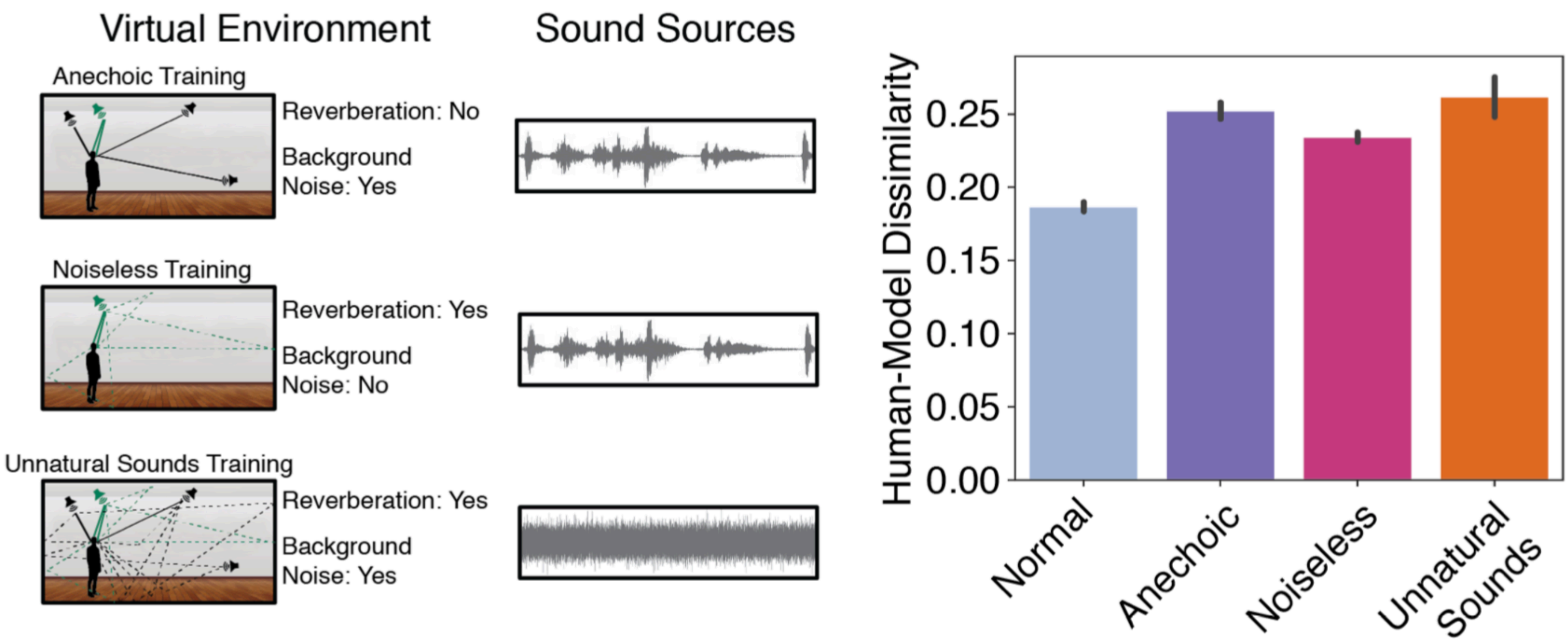




Mark Saddler Ray Gonzalez

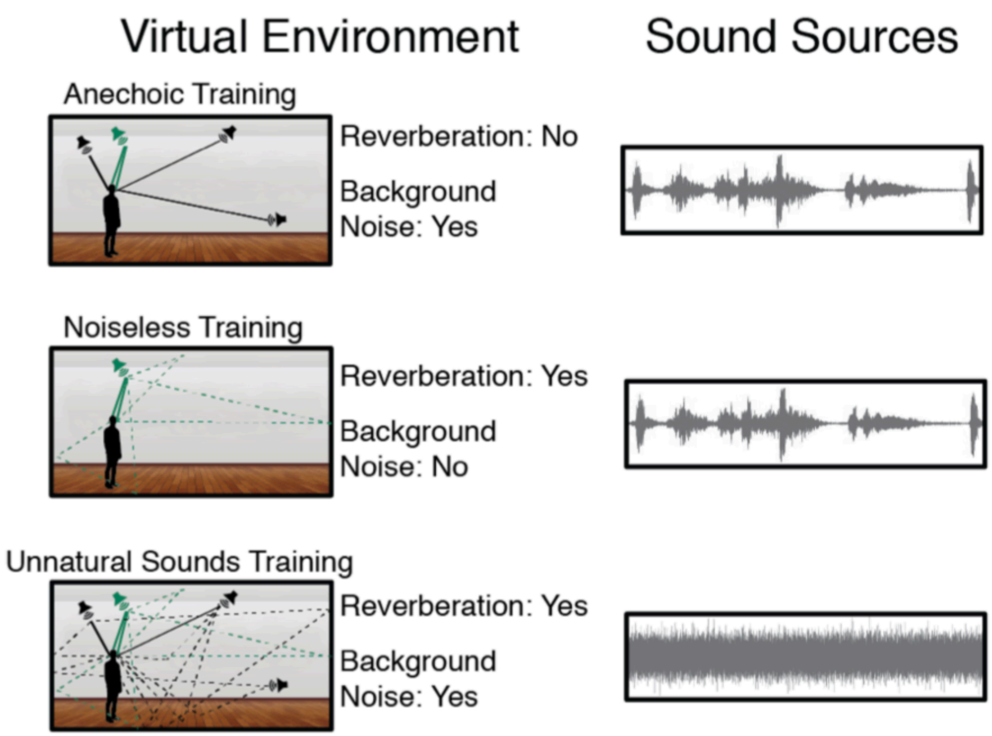
Similar result for sound localization: Model only resembles humans if optimized for natural conditions.

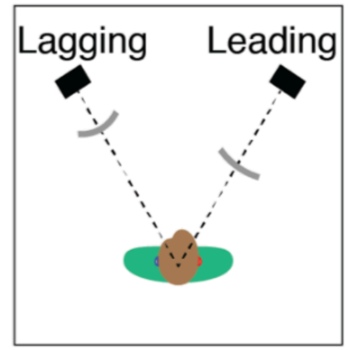
Alterations to training environment

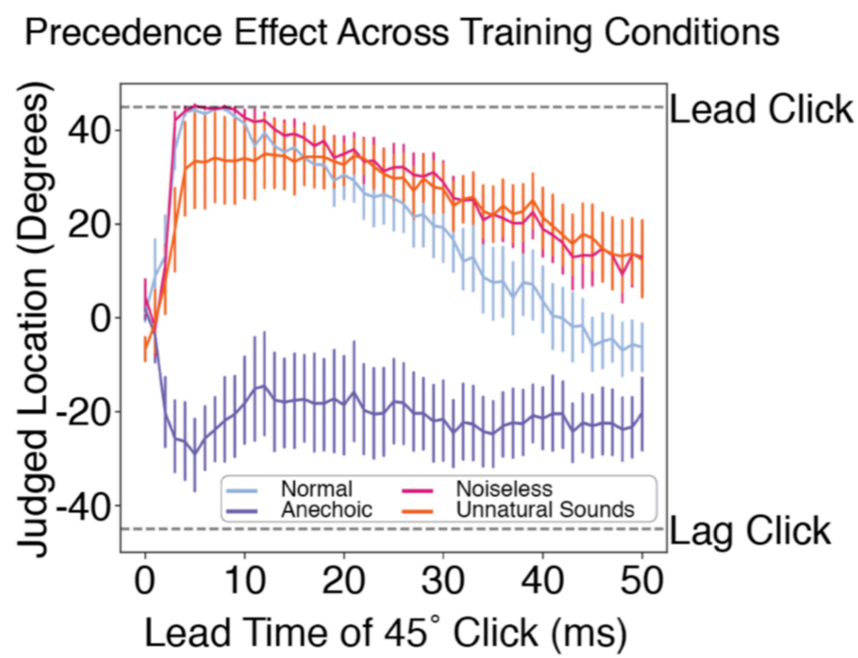


Example: precedence effect disappears selectively under anechoic training conditions

Alterations to training environment



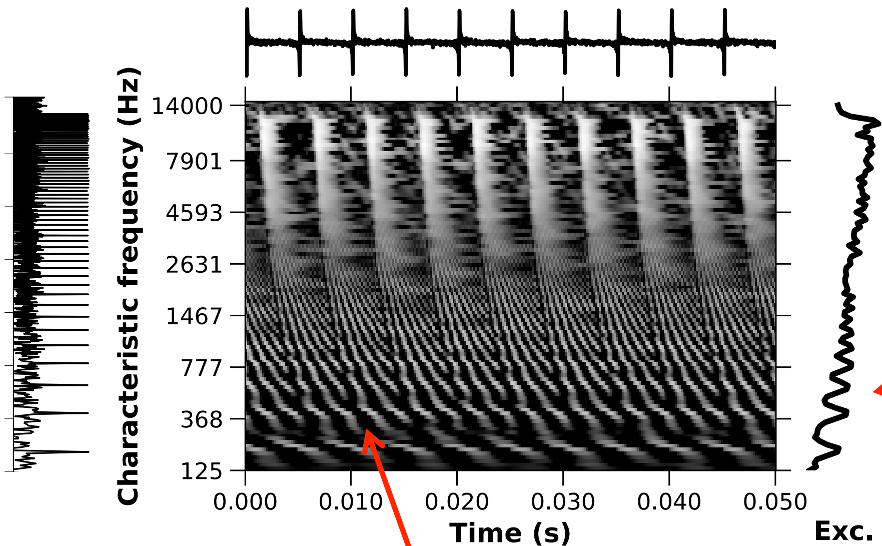




Trained neural networks can reveal performance characteristics of task-optimized mechanisms.

Conceptually similar to ideal observer models, but applicable to domains where deriving an ideal observer is intractable.

Longstanding controversy over timing vs. "place" information



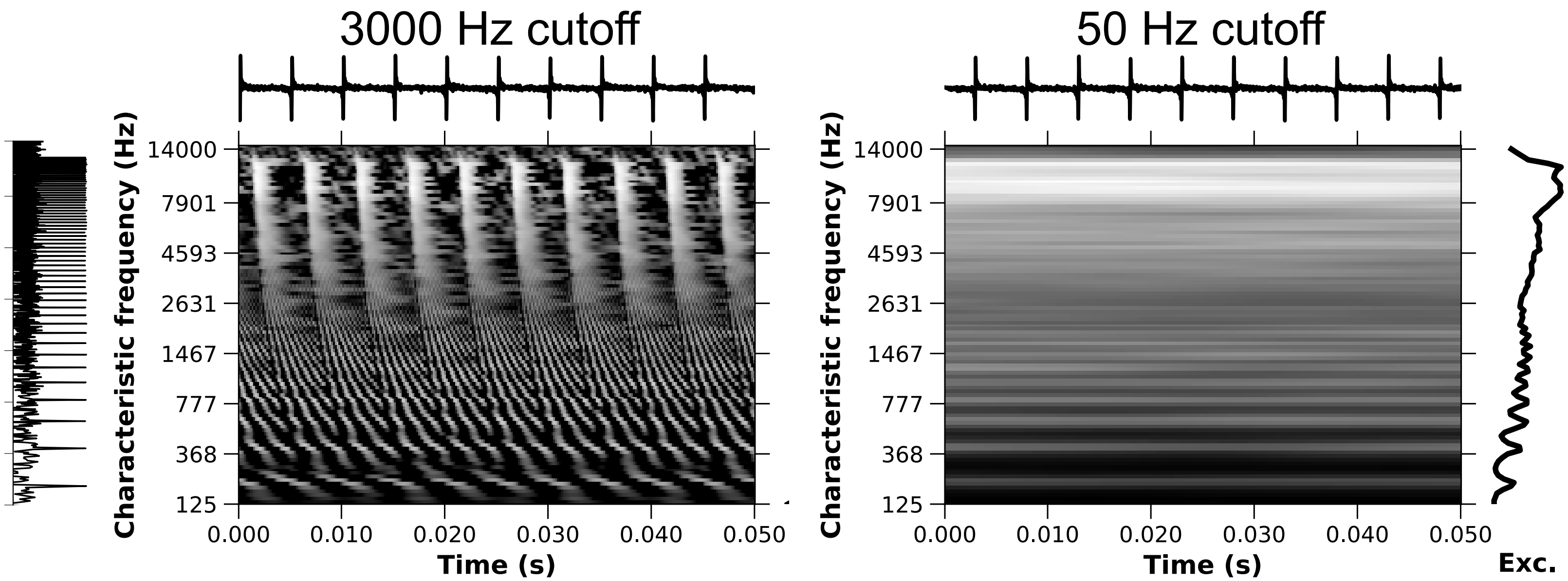
Excitation varies with place along cochlea, mirroring stimulus frequencies

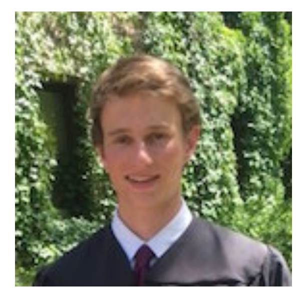
Membrane potential fluctuates with stimulus, up to ~4kHz

Mark Saddler

Longstanding controversy over timing vs. “place” information

Test by varying time constant of hair cell potential in cochlear model, retraining

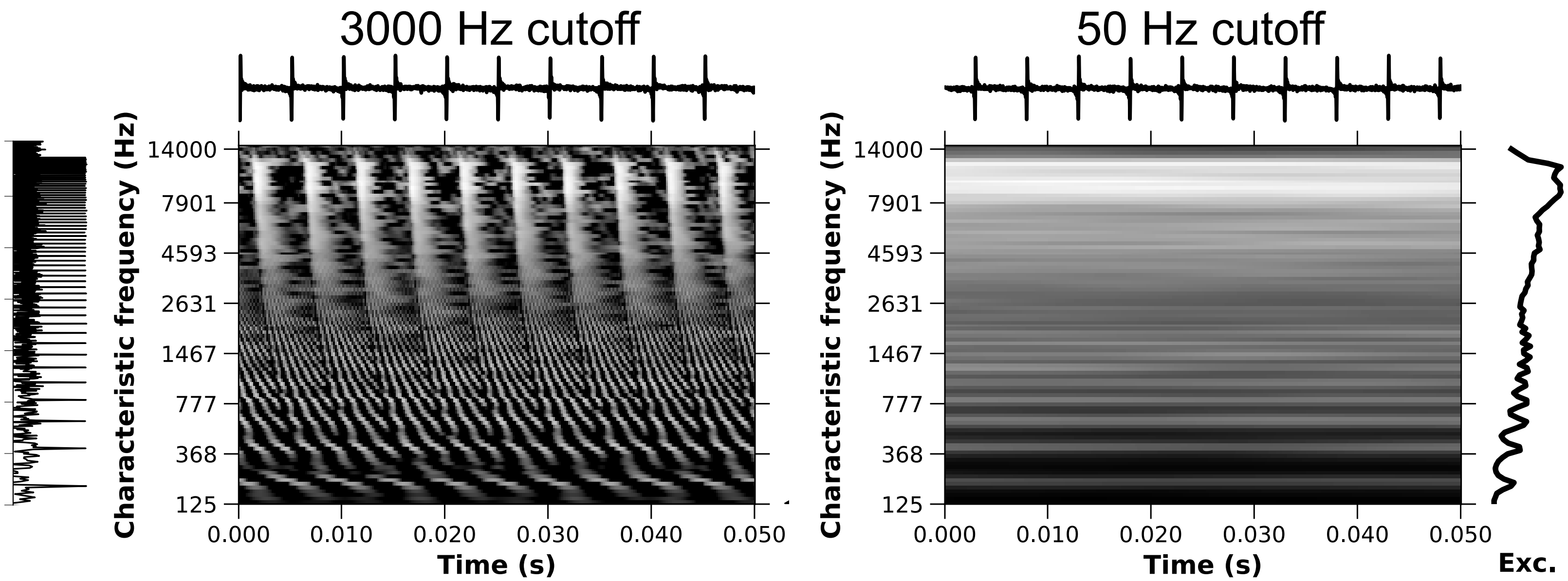


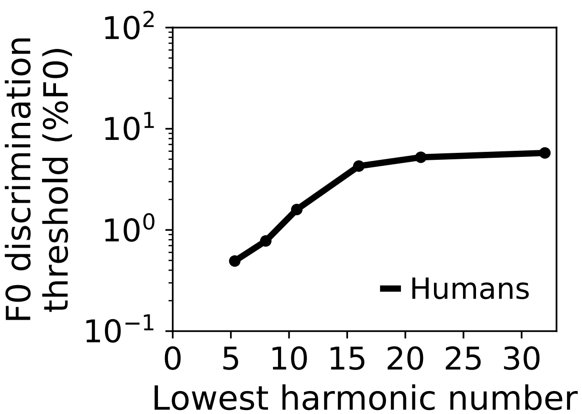


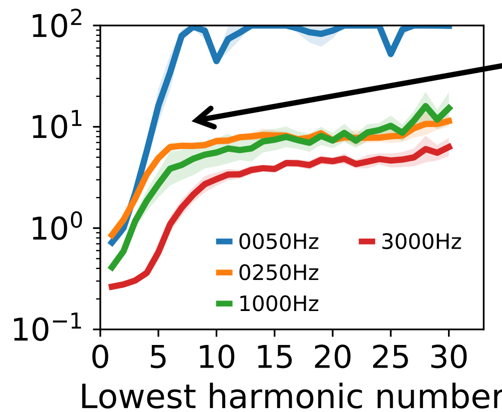
Mark Saddler

Longstanding controversy over timing vs. “place” information

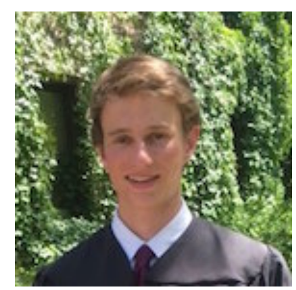
Test by varying time constant of hair cell potential in cochlear model, retraining







Inhuman performance if temporal information is limited.



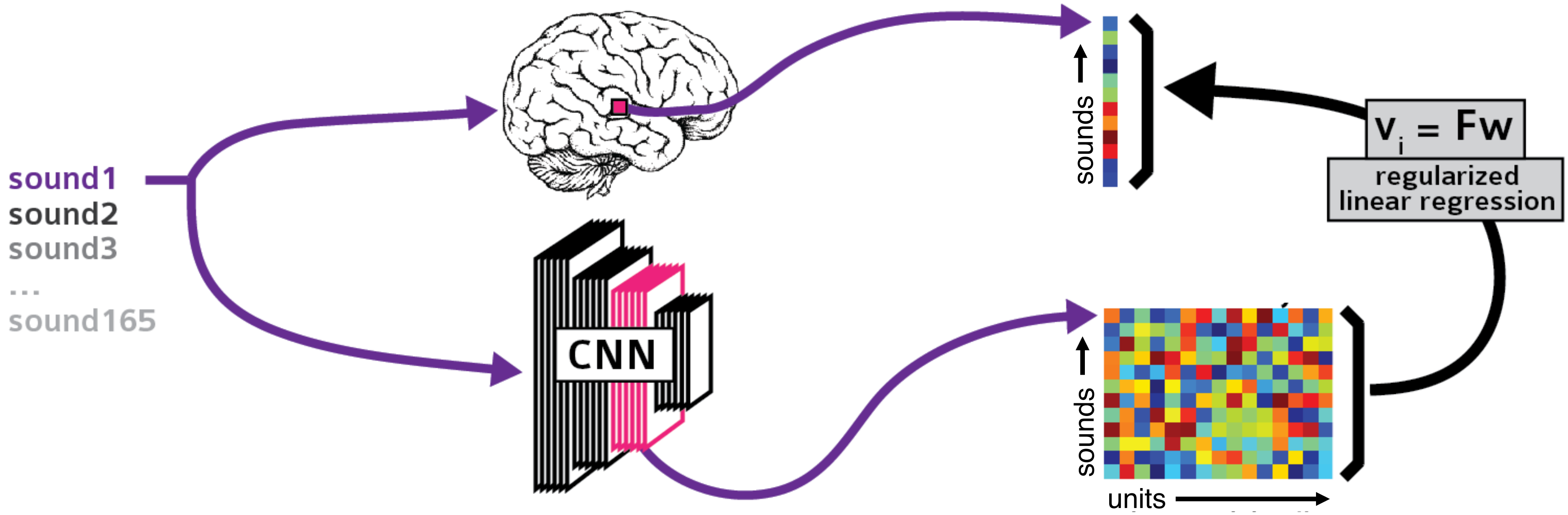
Mark Saddler

Trained neural networks exhibit similar performance characteristics to humans…

They also explain responses in the auditory cortex better than previous models.

Using learned features as encoding model

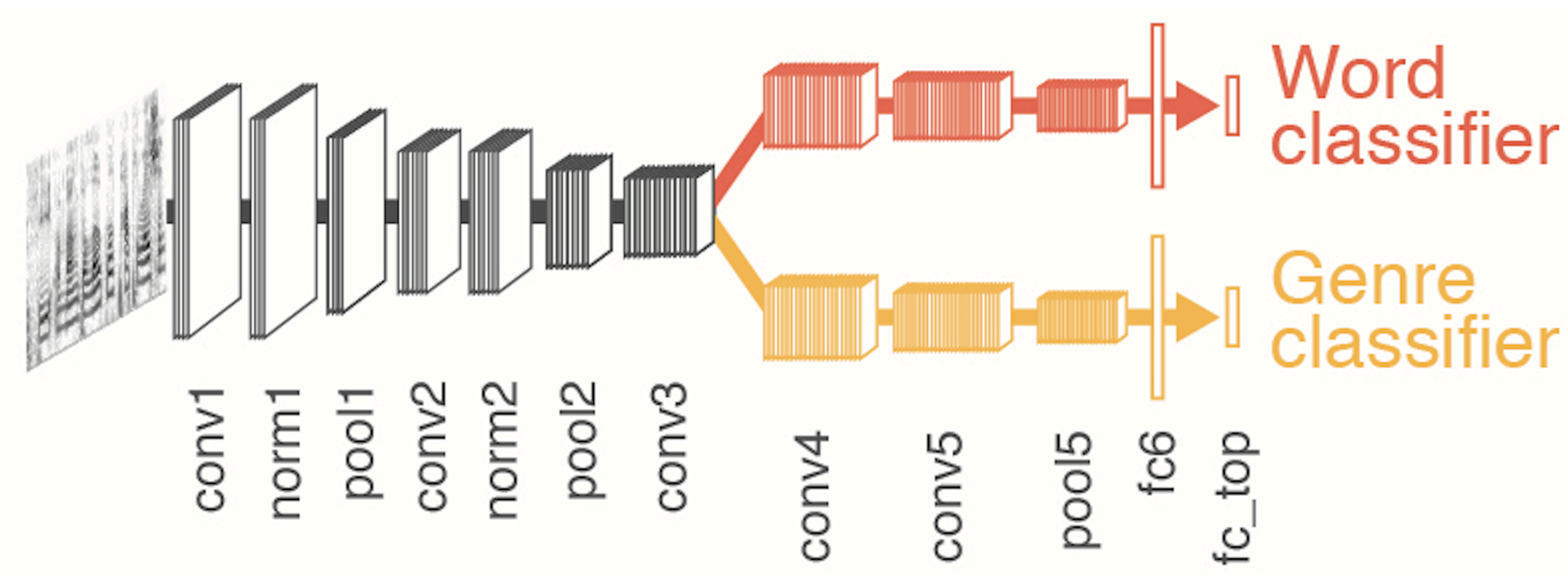
Each voxel = weighted sum of time-averaged unit responses in a given layer



Cross-validated regularized linear regression to predict voxel's response

Best current model: dual pathways

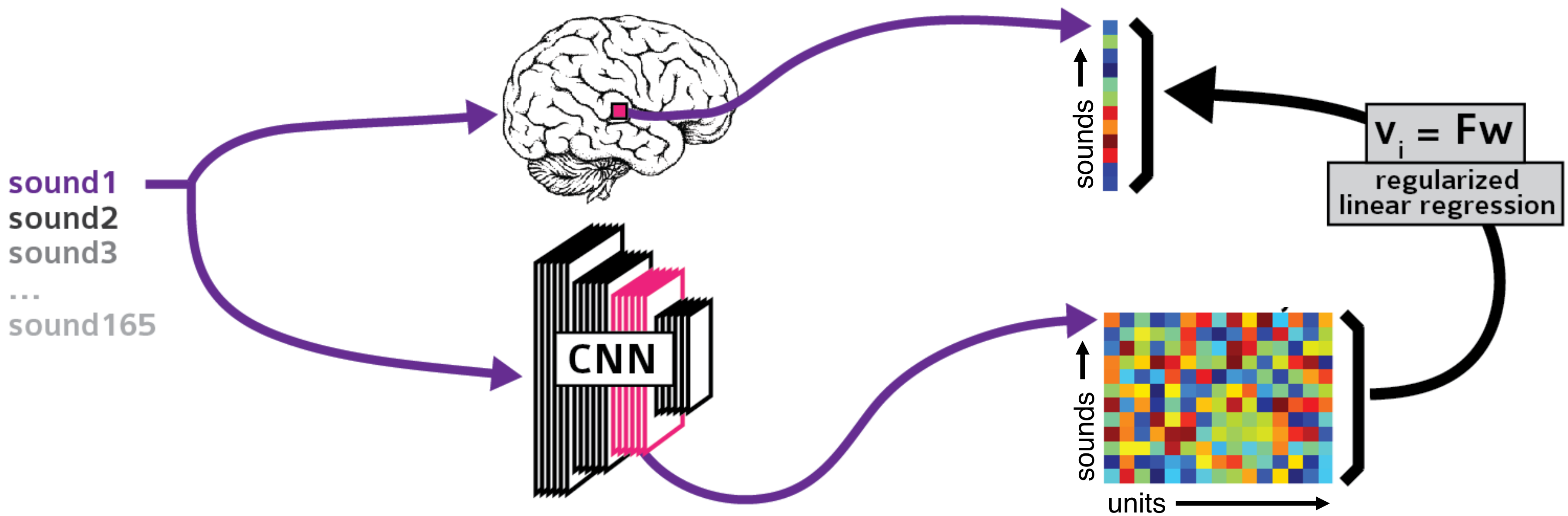
- Optimizing across architectures yields split between speech and music.
- Speech and music share early stages of computation



Kell et al., Neuron, 2018

Using learned features as encoding model

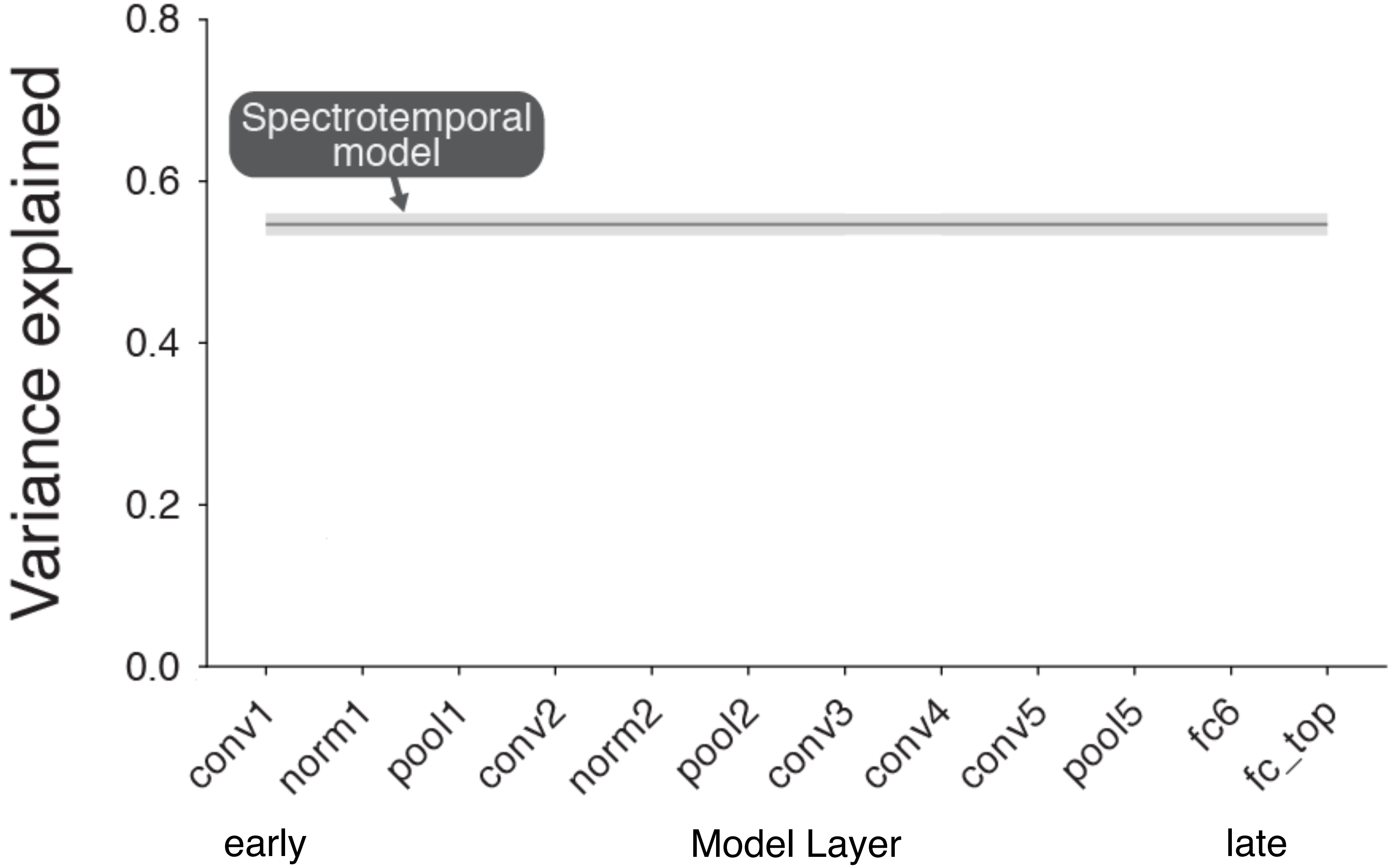
Each voxel = weighted sum of time-averaged unit responses in a given layer



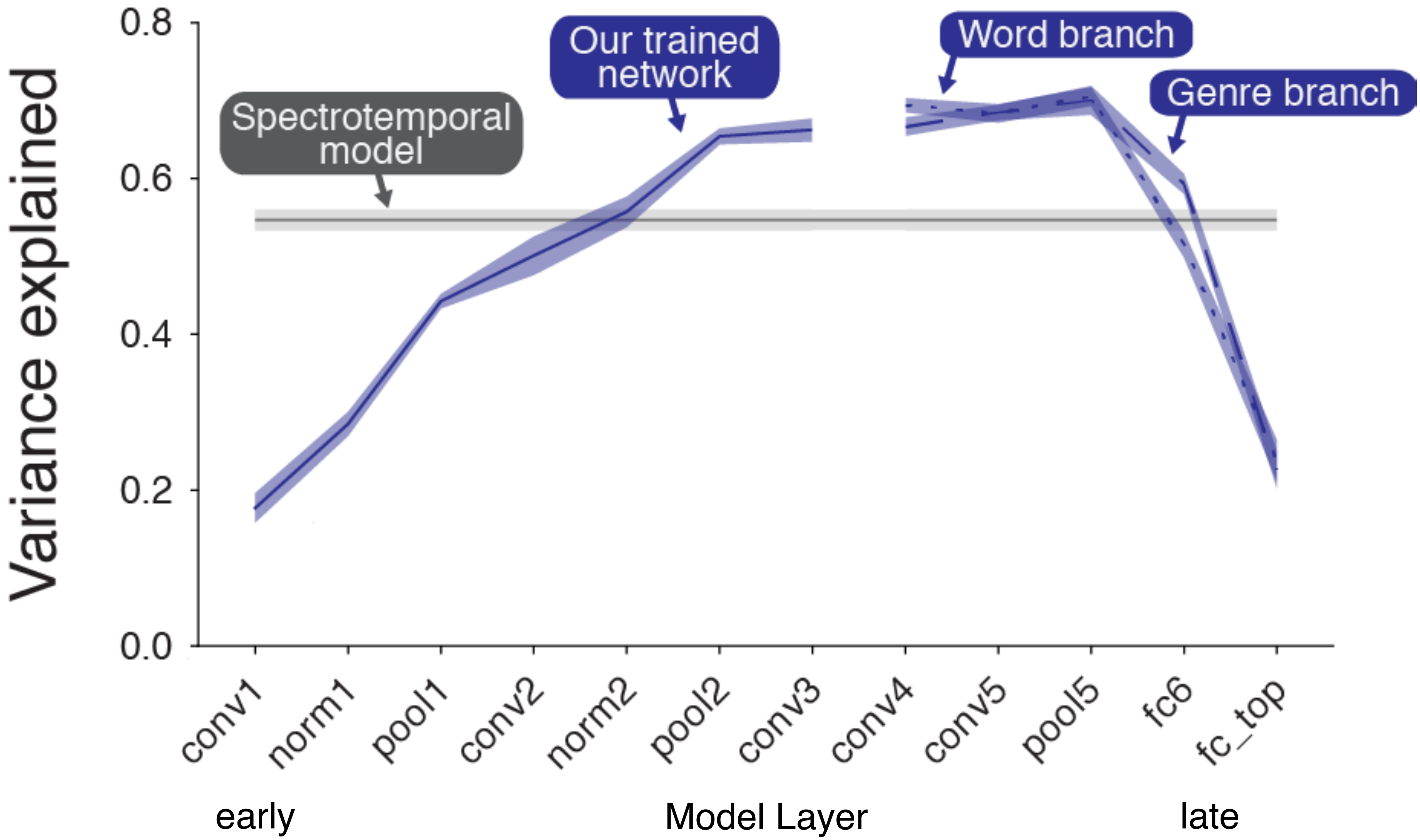
Cross-validated regularized linear regression to predict voxel's response

Baseline:
Identical procedure with the spectrotemporal filter model

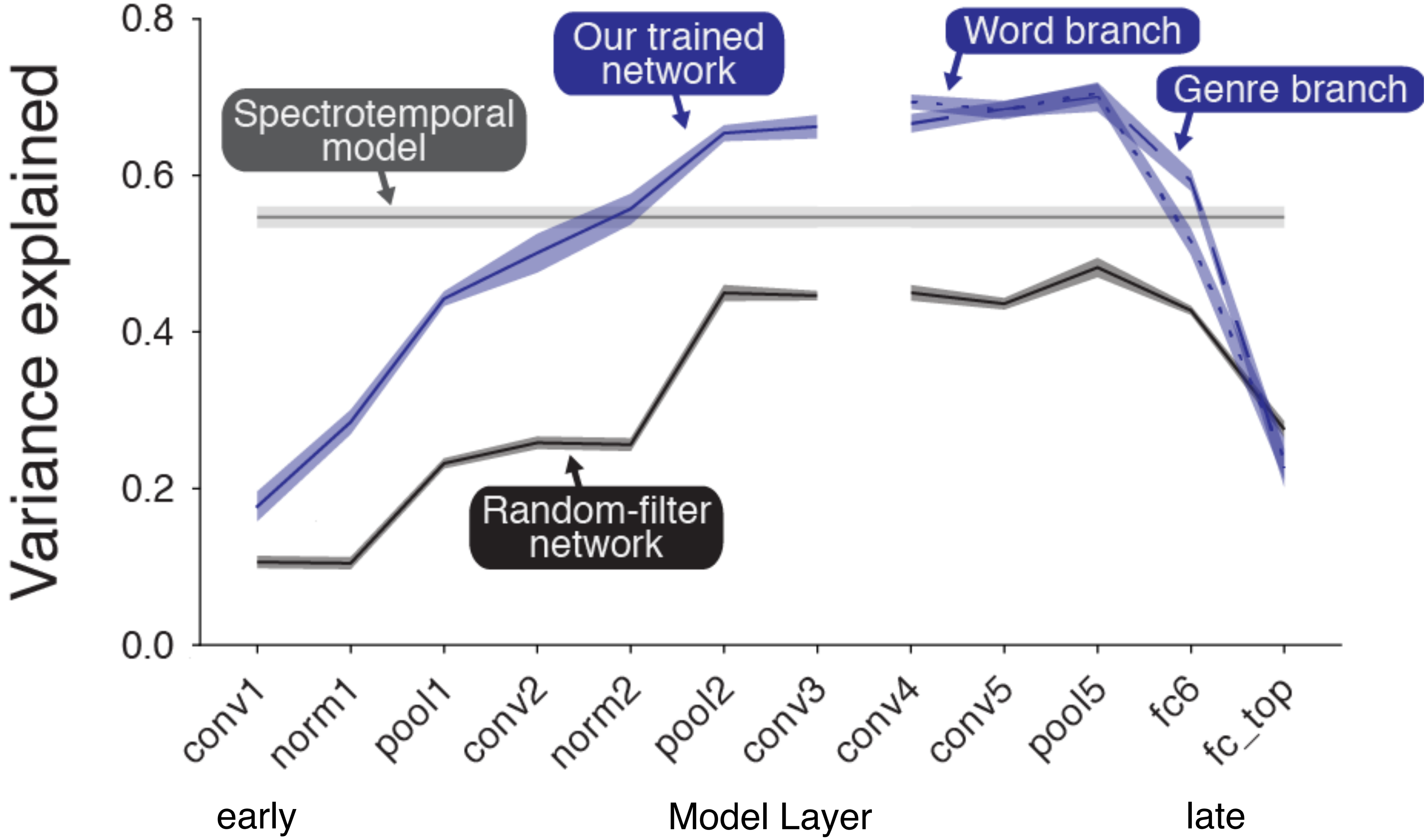
Median variance explained across all of auditory cortex:

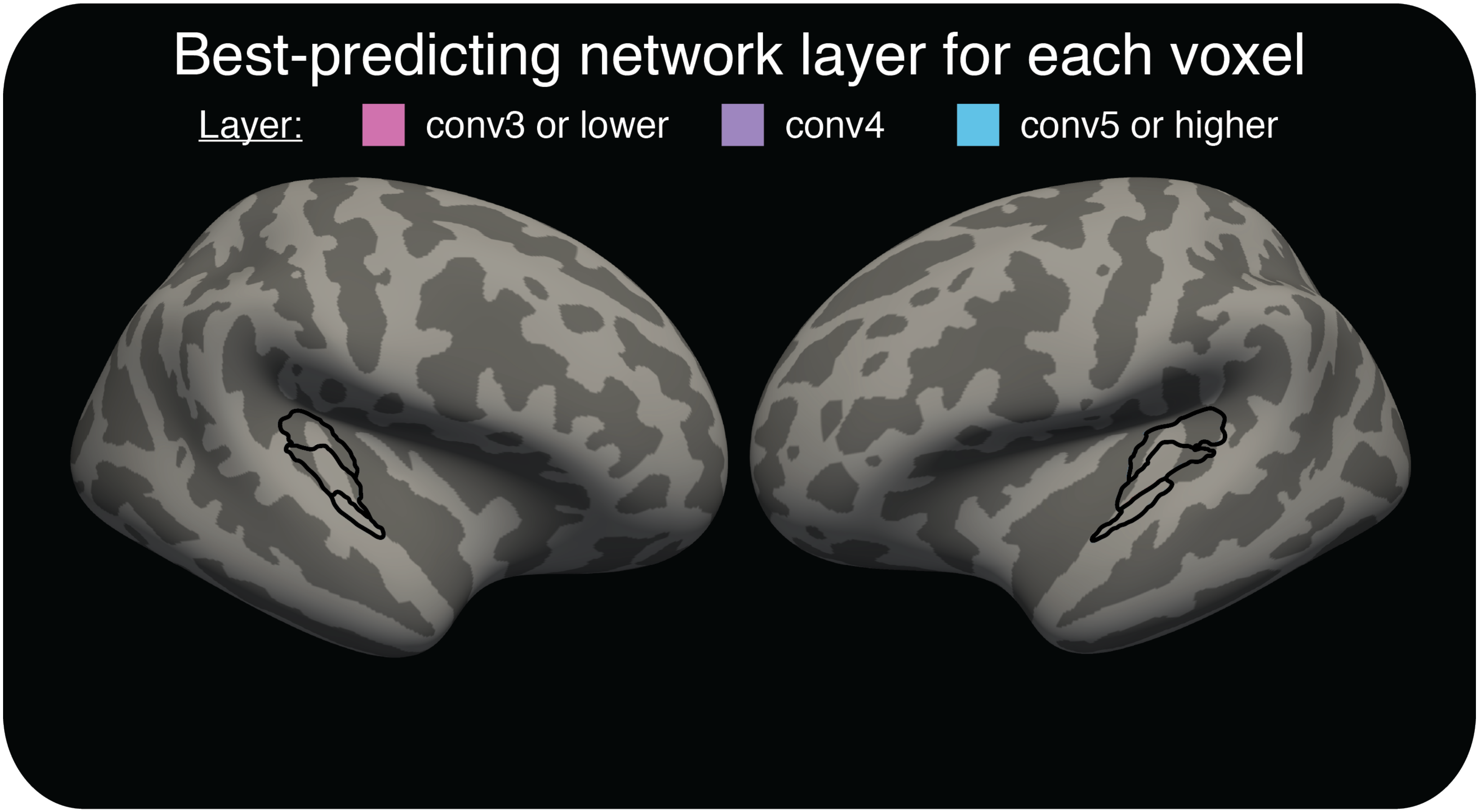


Middle layers of model best predict cortical voxel responses



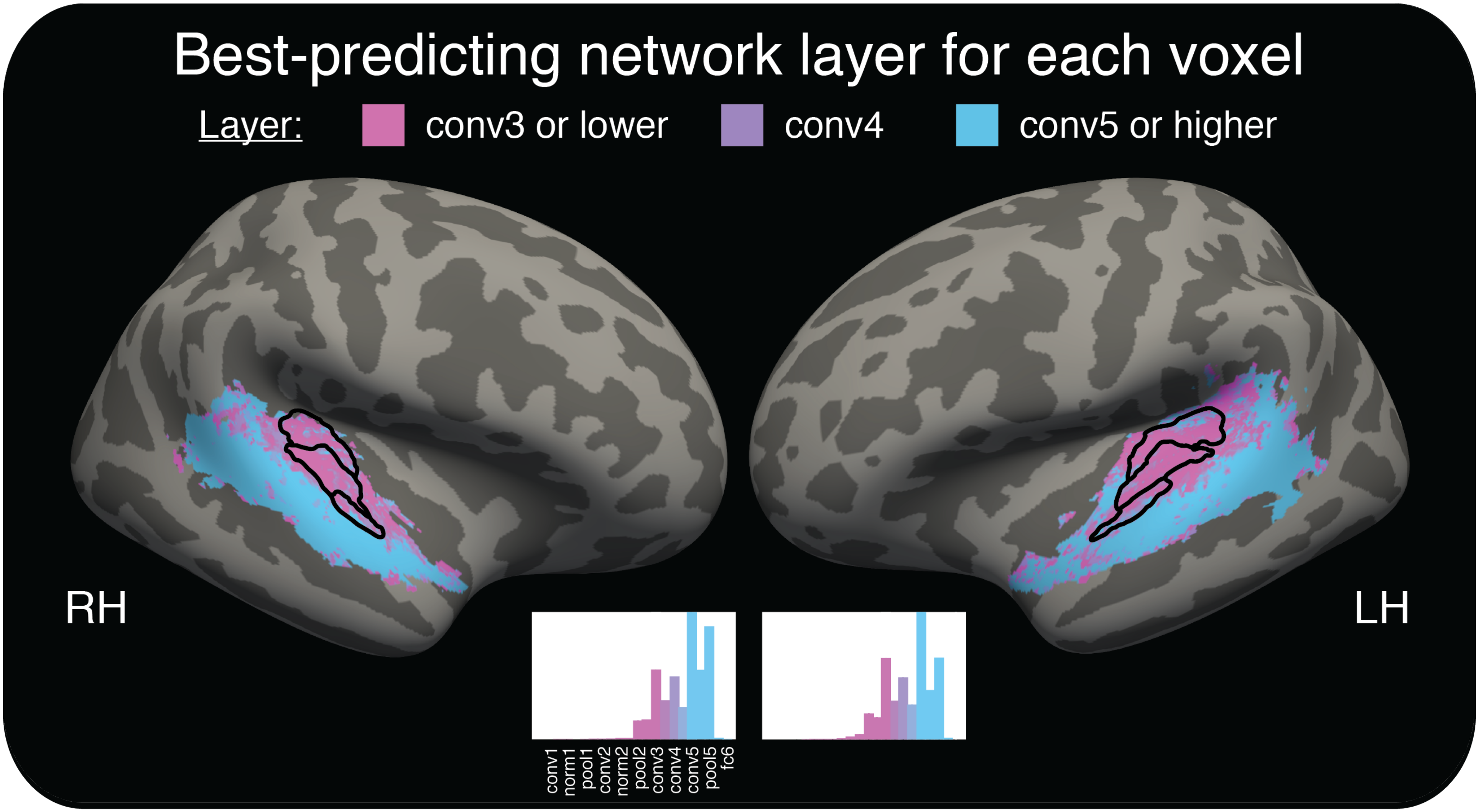
Middle layers of model best predict cortical voxel responses





Kell et al., Neuron, 2018

Suggestive of hierarchical organization of human auditory cortex



Pretty clear evidence of two stages (core/belt)
But not obvious tertiary structure.

Take-Home Messages, Part 1

After training on natural auditory tasks with natural sounds:

- Pretty good matches to human behavioral experiments
  - Speech recognition in noise
  - Sound localization
  - Pitch perception

- Best current predictions of auditory cortical responses

Manipulation of training conditions shows that similarity is a function of optimization for natural tasks/sounds, cochlea

  - Provides insight into origins of human behavioral traits

Degrading simulated cochlear input to the neural network reproduces characteristics of human hearing impairment

Plan for Today

- Summary of recent successes of our neural network models of hearing
- Discussion of current model shortcomings

Take-Home Messages, Part 2

Metamers of neural networks provide a way to reveal model invariances

- Metamers of deep layers of standard neural network models are not metameric for humans
  - Not even recognizable to humans
  - True for vision and auditory networks

- Model metamers can be made more human-recognizable with some architectural modifications (reducing aliasing)

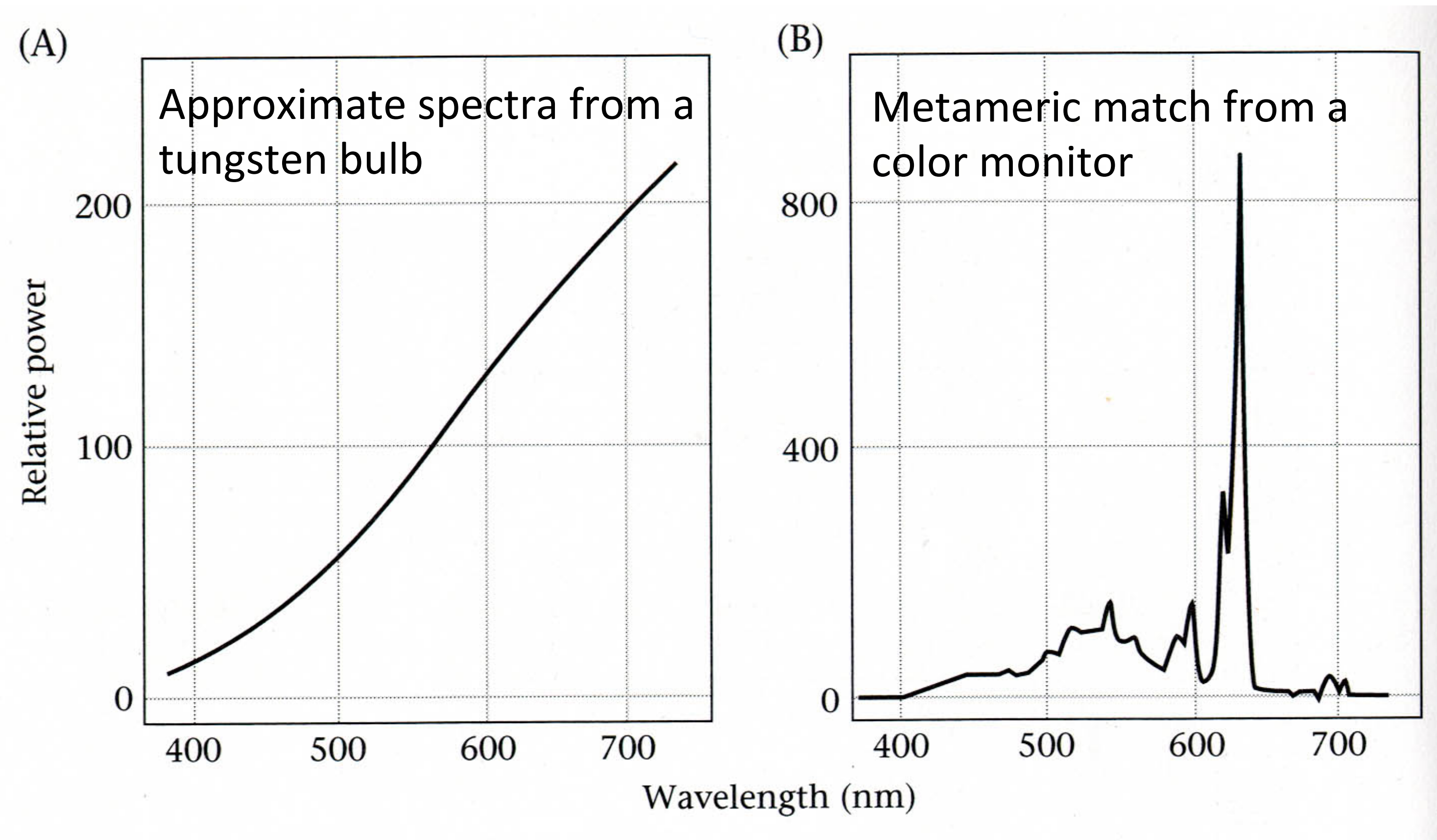
  - And by making models more robust to adversarial examples (for reasons we don't yet fully understand)

  - But divergences remain

Can network invariances be revealed with model metamers?

Metamers – physically distinct stimuli that are indistinguishable to observer

Classic example: color vision



Can network invariances be revealed with model metamers?

Metamers – physically distinct stimuli that are indistinguishable to observer

Classic example: color vision

$$\begin{bmatrix} \text{L} \\ \text{M} \\ \text{S} \end{bmatrix} = \begin{pmatrix} \text{Spectral sensitivity of L photopigment} \\ \text{Spectral sensitivity of M photopigment} \\ \text{Spectral sensitivity of S photopigment} \end{pmatrix} \text{x} \begin{bmatrix} \text{Light projected into eye} \end{bmatrix}$$

But also evident in human texture perception, crowding

cf Julesz, Rosenholtz, Simoncelli

Can network invariances be revealed with model metamers?

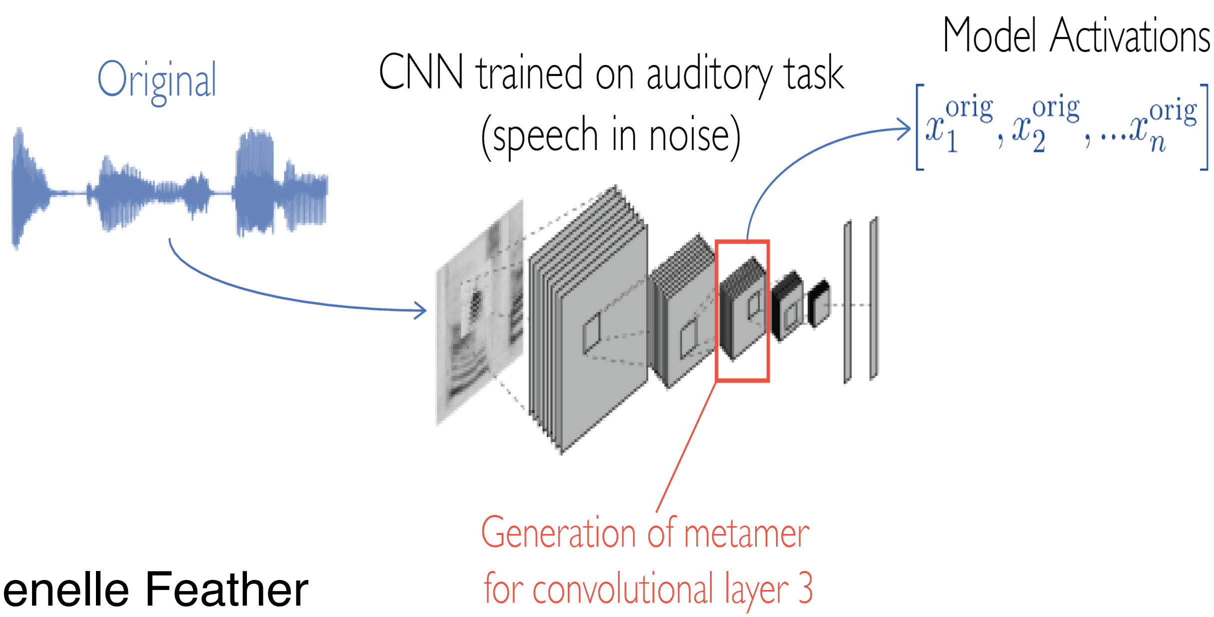
Metamers – physically distinct stimuli that are indistinguishable to observer

Instantiation of invariant recognition within network should produce model metamers

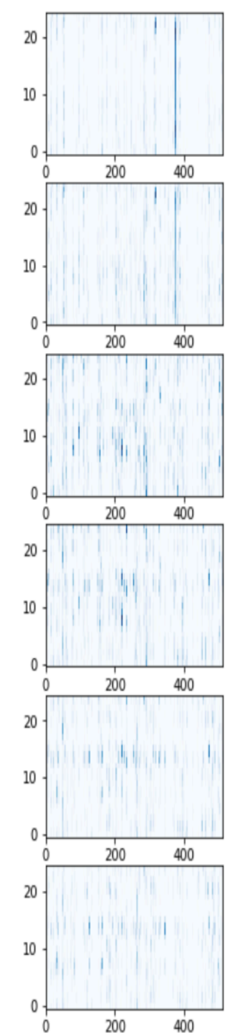
-could reveal learned transformations

-could provide another test of whether model captures human perception

Network invariances can be revealed with model metamers

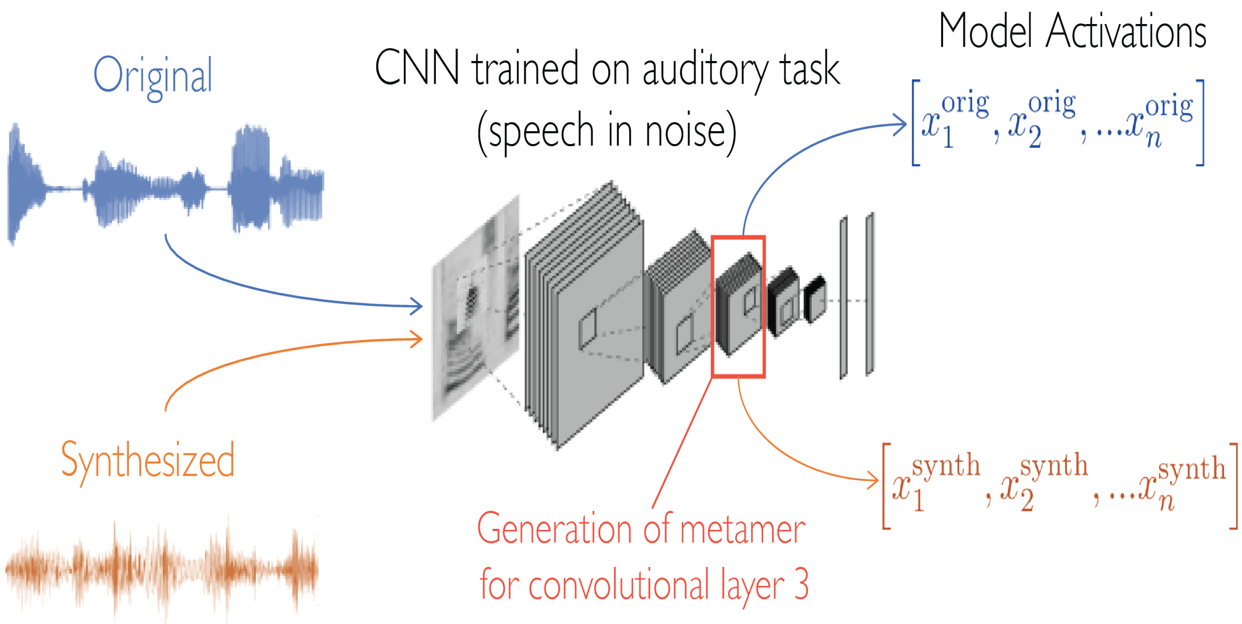


Original

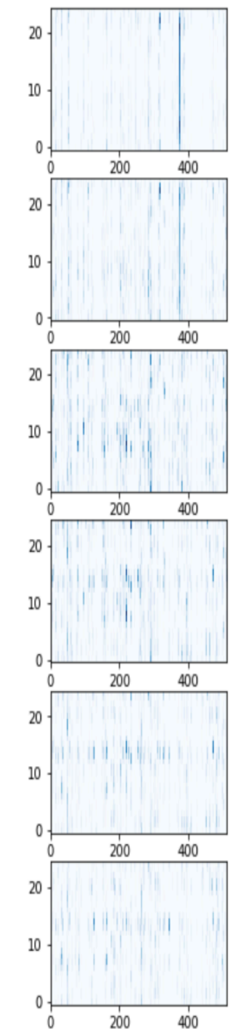


Jenelle Feather

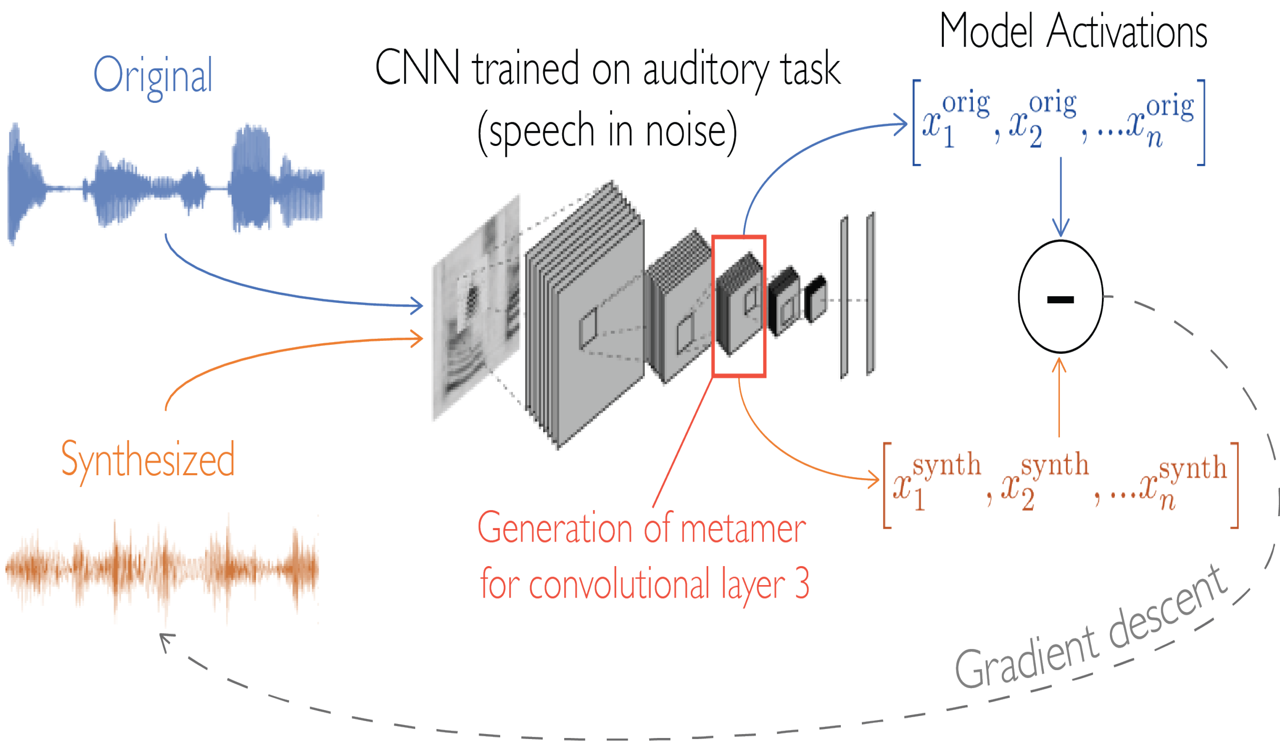
Network invariances can be revealed with model metamers

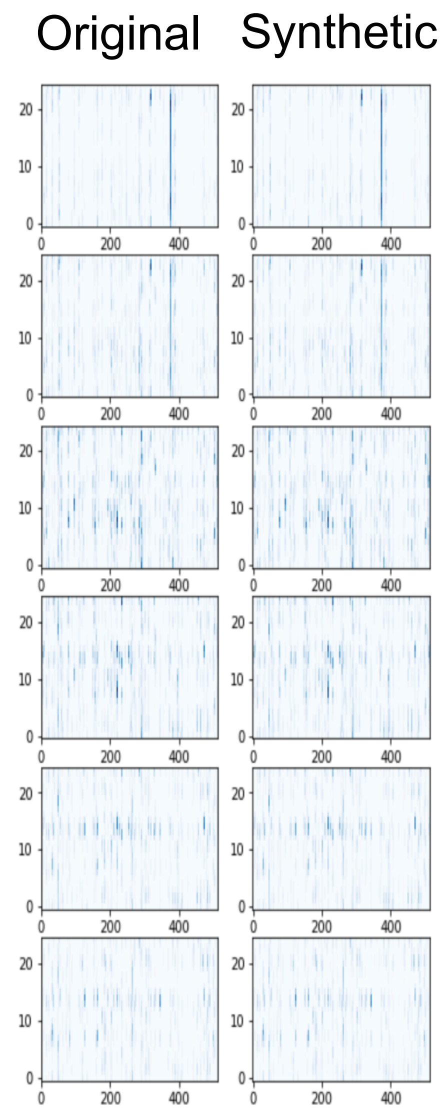


Original



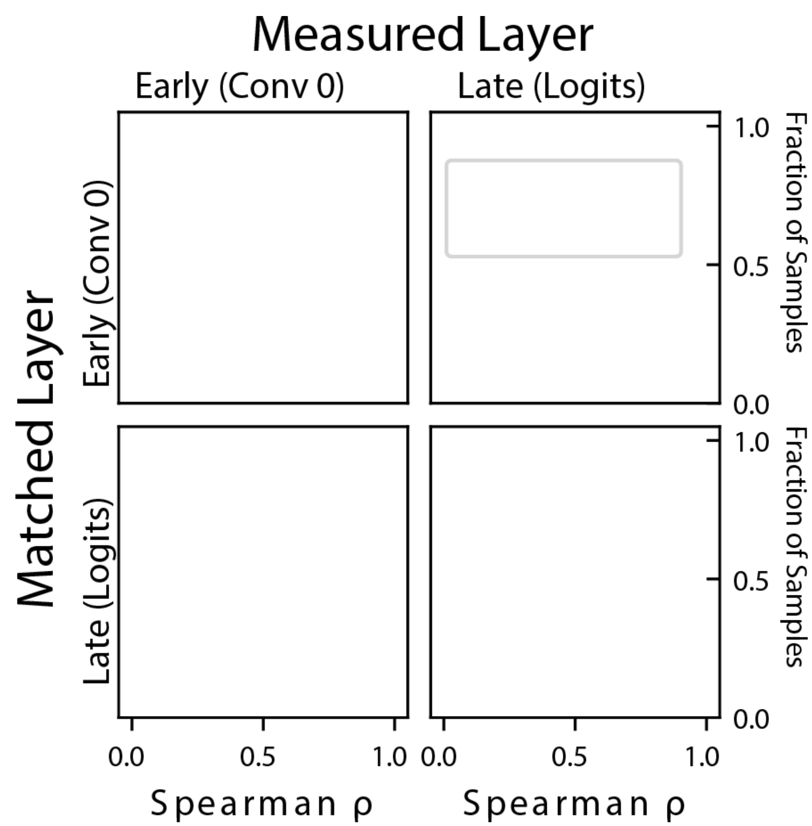
Network invariances can be revealed with model metamers





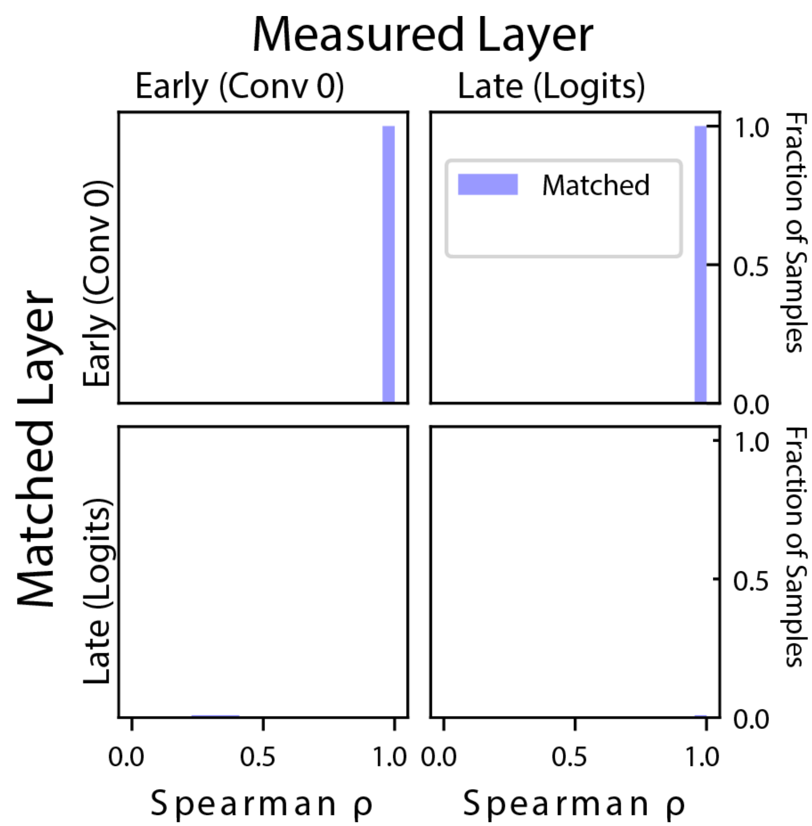
- Network's response within a layer is matched
- All subsequent layers are also matched.
- Decision about stimulus is thus the same.

Network invariances can be revealed with model metamers



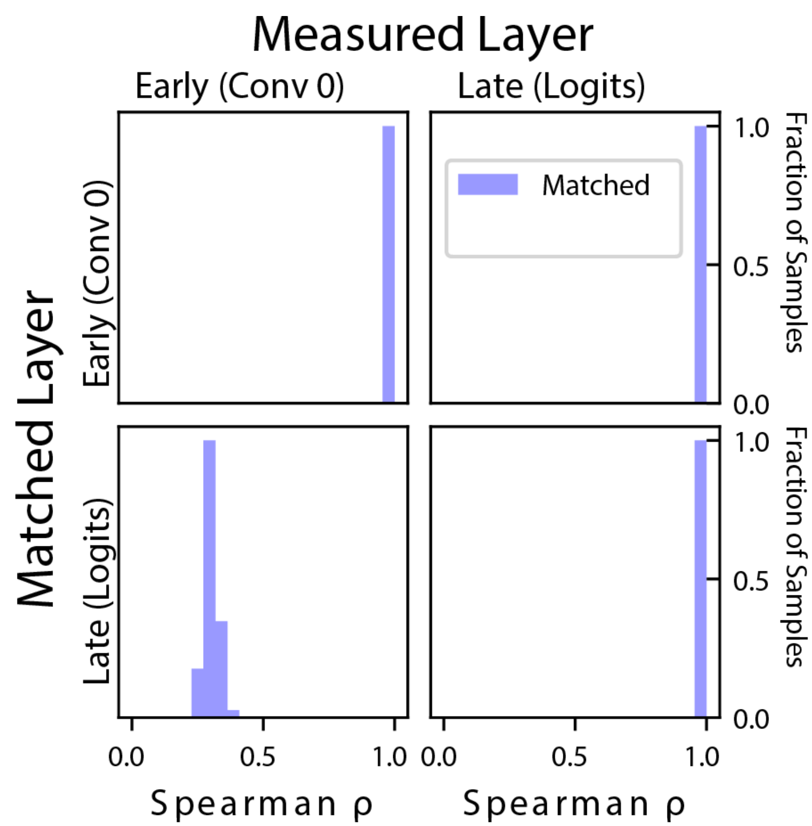
- Network's response within a layer is matched
- All subsequent layers are also matched.
- Decision about stimulus is thus the same.

Network invariances can be revealed with model metamers



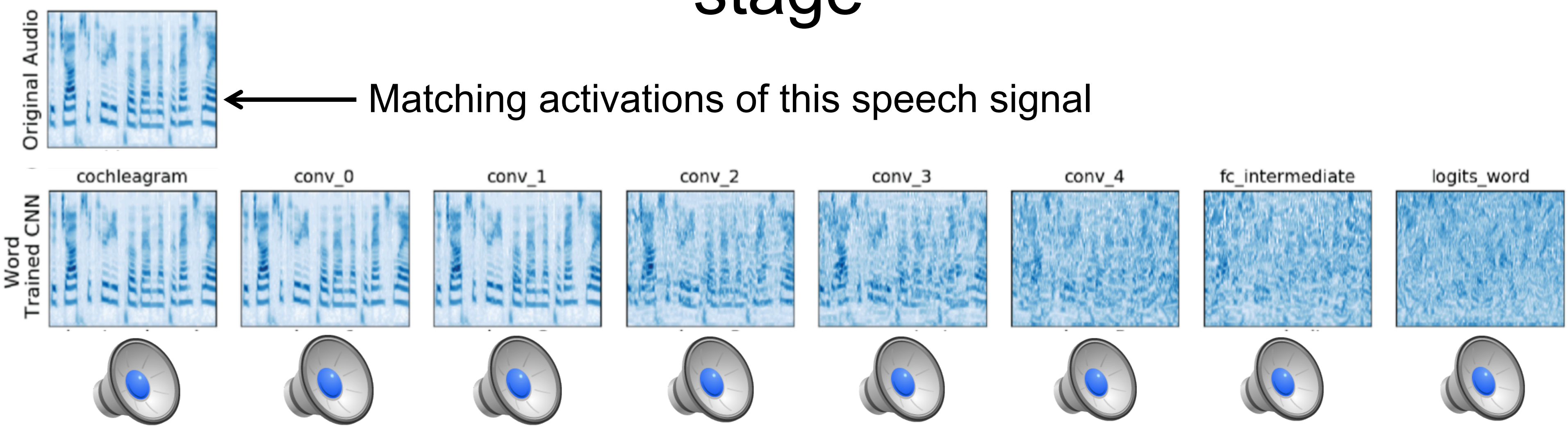
- Network's response within a layer is matched
- All subsequent layers are also matched.
- Decision about stimulus is thus the same.

Network invariances can be revealed with model metamers



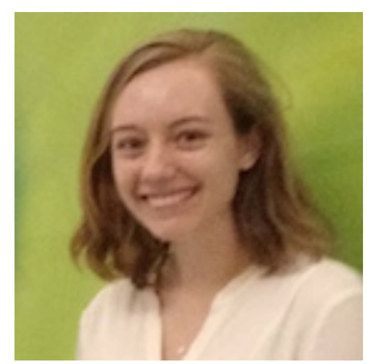
- Network's response within a layer is matched
- All subsequent layers are also matched (but not earlier).
- Decision about stimulus is thus the same.

Example metamers from each convolutional stage



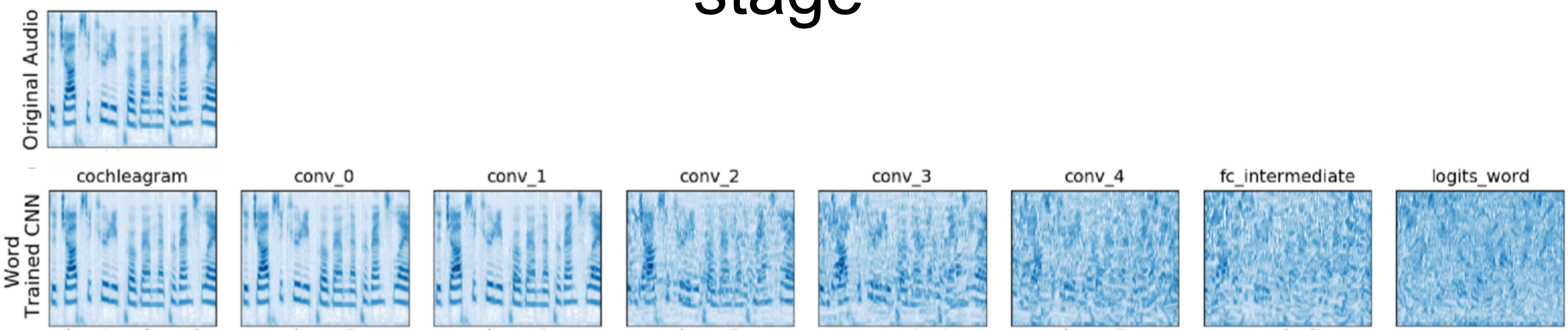
- Metamers are fully recognizable to network (by design), but become progressively unintelligible to humans
- Evaluate with recognition task (more conservative than a test of human metamerism)

Jenelle Feather

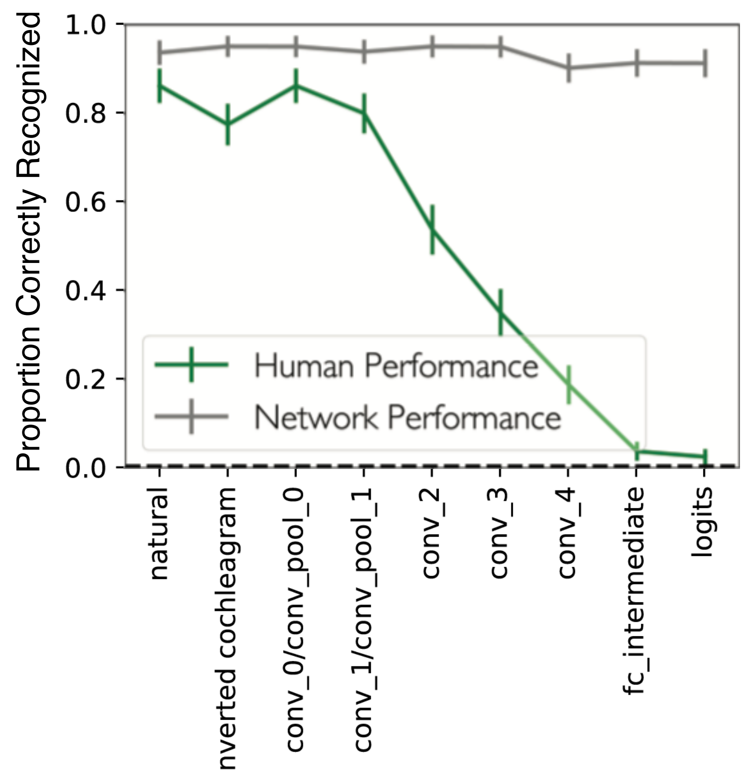


Feather et al., NeurIPS, 2019

Example metamers from each convolutional stage

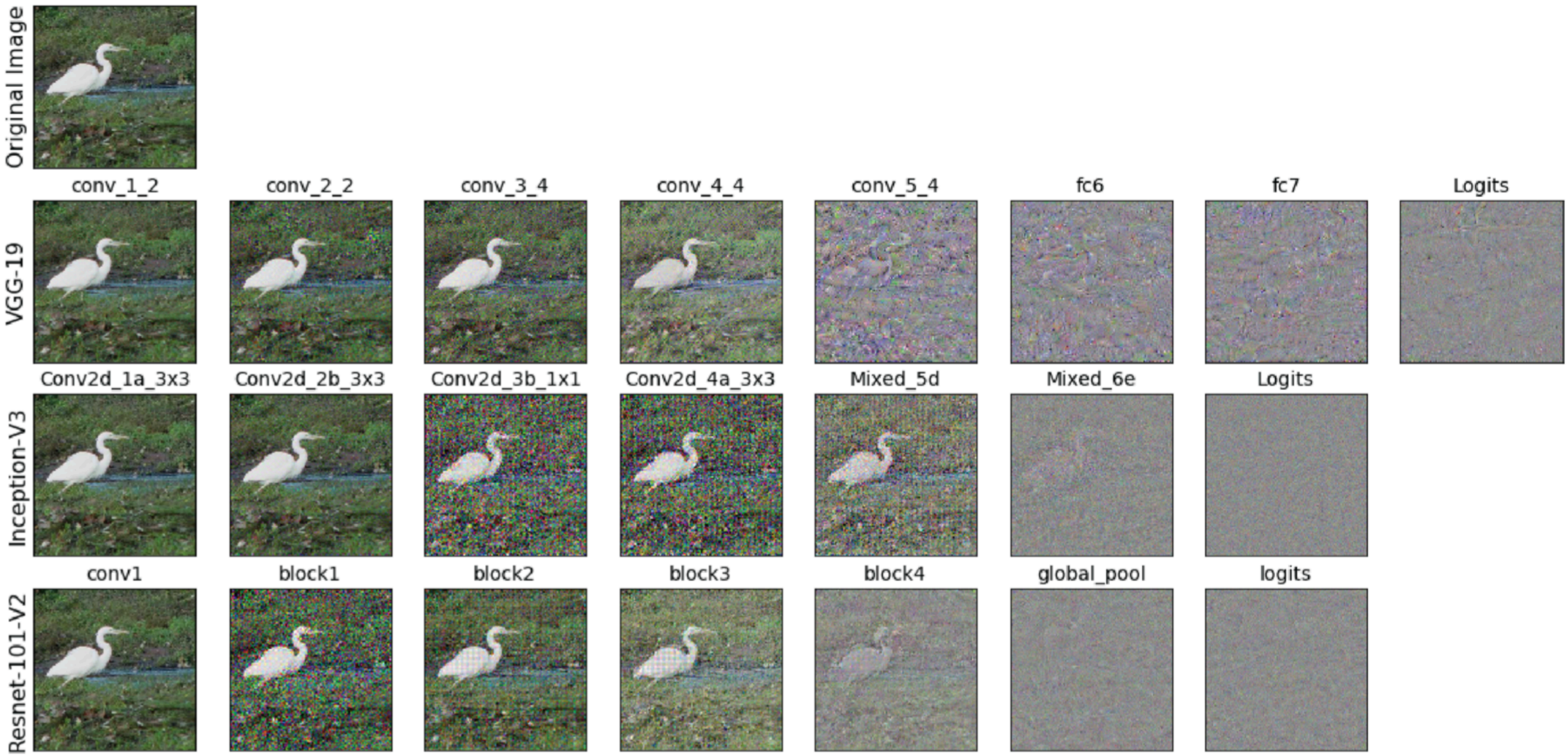


- Metamers are fully recognizable to network (by design), but become progressively unintelligible to humans
- Evaluate with recognition task (more conservative than a test of human metamerism)



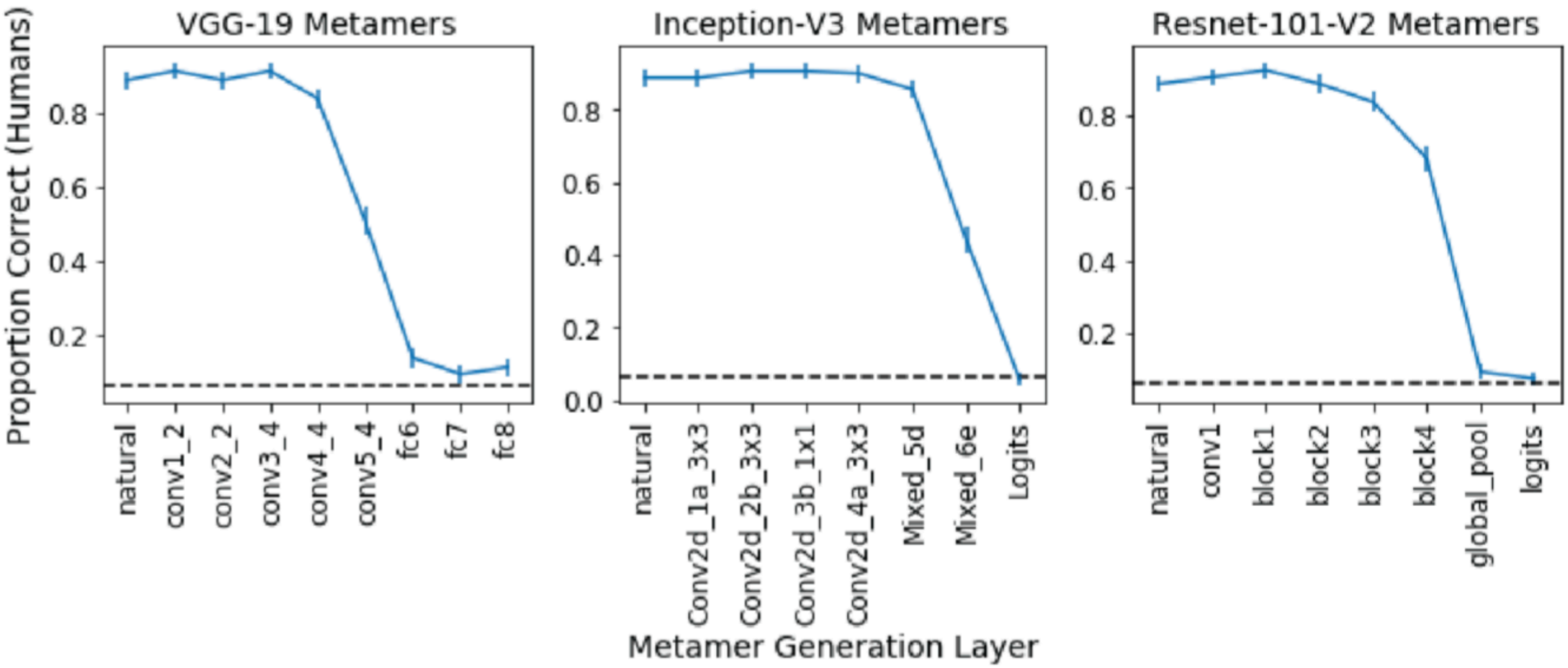
Feather et al., NeurIPS, 2019

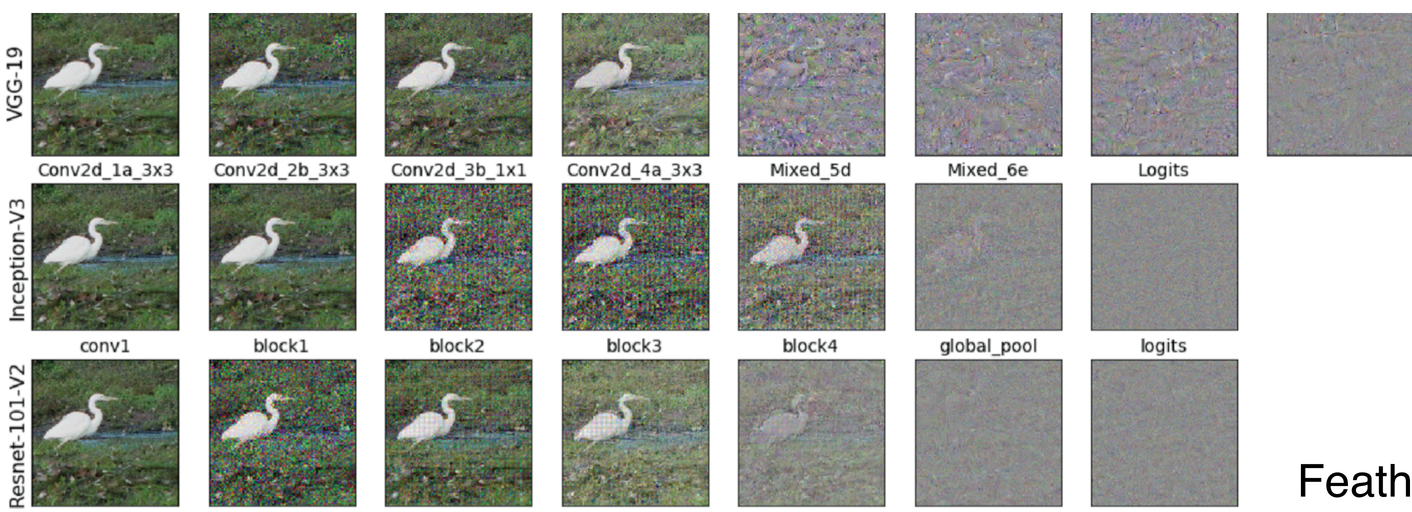
Qualitatively similar results for vision networks:



cf Mahendran and Vedaldi, 2015

Qualitatively similar results for vision networks:

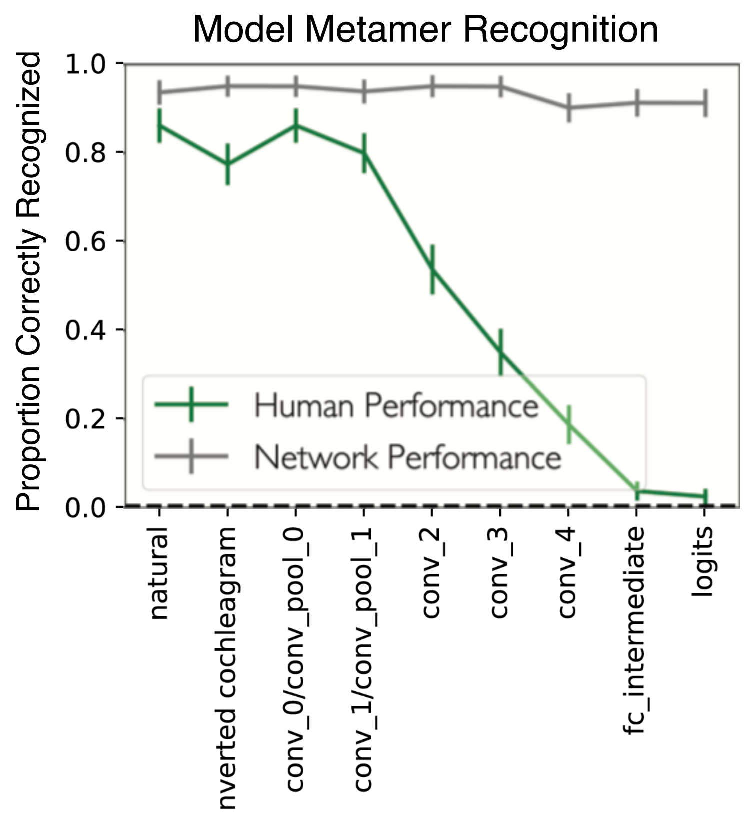




Feather et al., NeurIPS, 2019

Model metamers are often unrecognizable to humans

- In contrast to similar behavior with natural sounds, divergent behavior with unnatural signals
- Substantial inconsistency with biological perceptual systems
- Strong benchmark for evaluating sensory models



Jenelle Feather

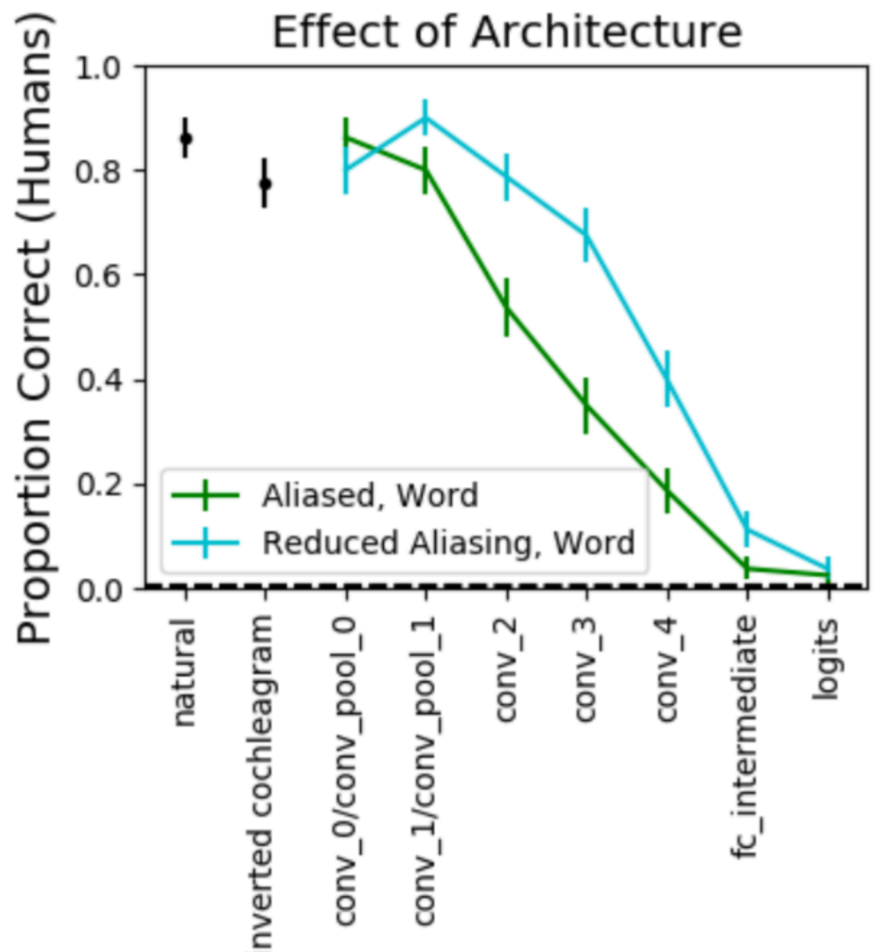
Feather et al., NeurIPS, 2019

Reasons for pessimism?

- Many functions are consistent with the training data
- Most guarantees of “reasonable” behavior only hold within training distribution
- Perhaps divergent metamers are expected and unavoidable?

Reasons for optimism?

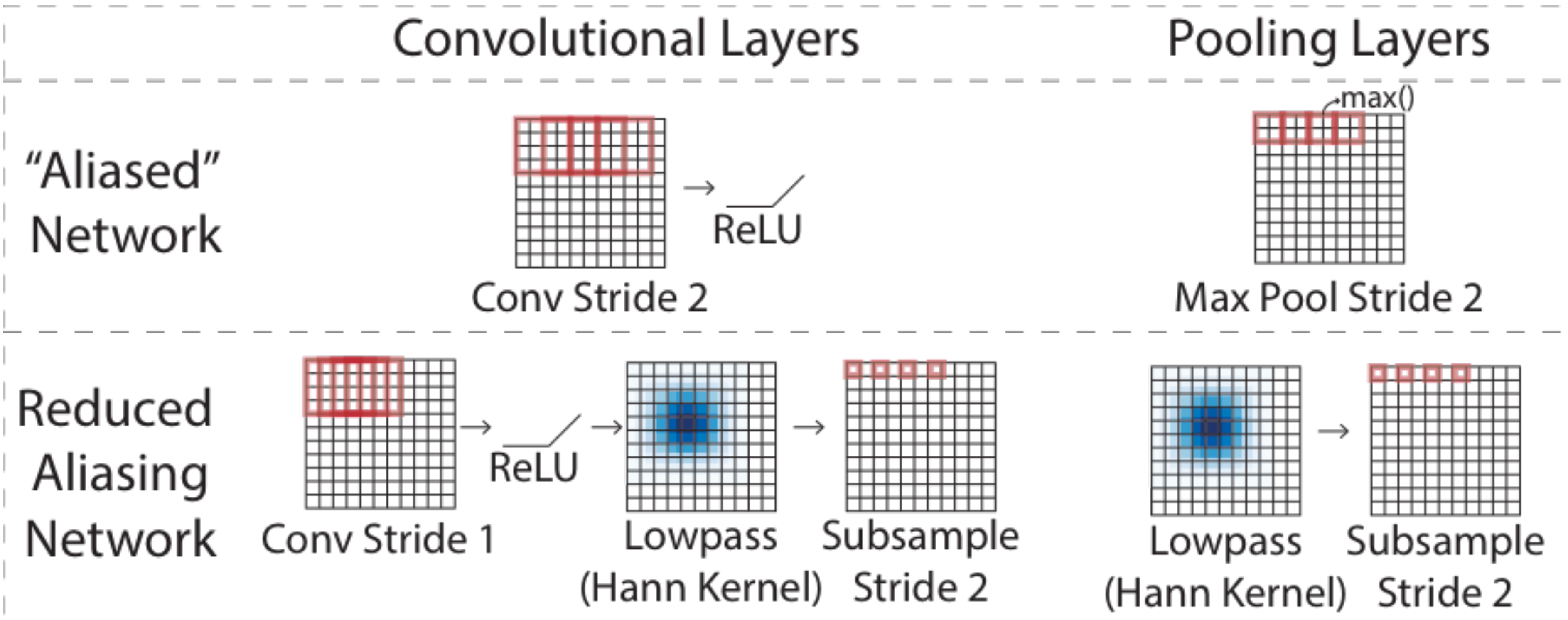
- Reducing aliasing improves human-recognizability of model metamers
- Consistent with classical signal processing intuitions about biological sensory systems



cf Zhang 2019; Azulay and Weiss, 2018
Henaff and Simoncelli, 2015

Feather et al., NeurIPS, 2019

Reasons for optimism?

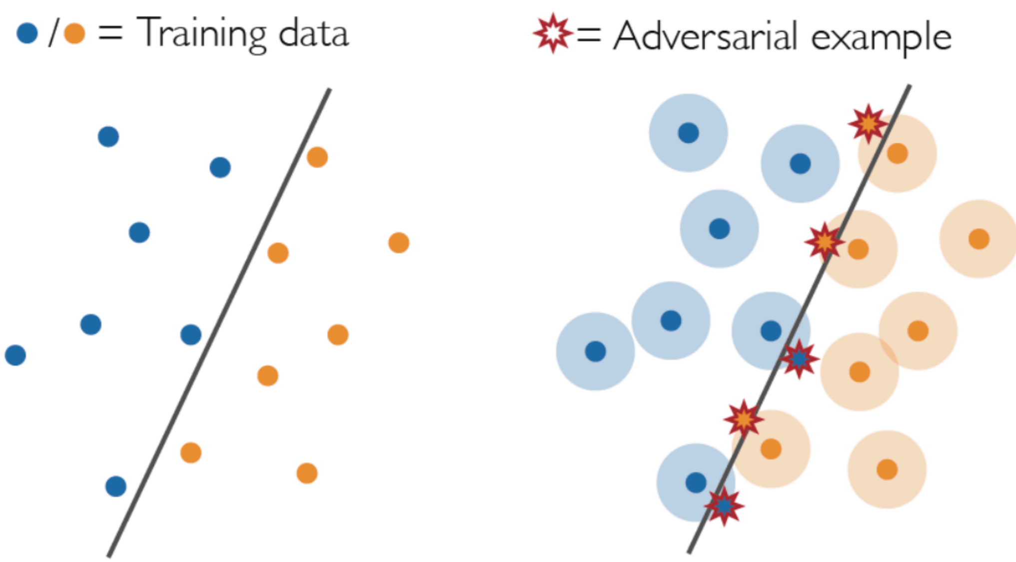


cf Zhang 2019; Azulay and Weiss, 2018
Henaff and Simoncelli, 2015

Feather et al., NeurIPS, 2019

How to address model inadequacies?

Other major divergence between neural networks and human perception: adversarial examples



Standard Training

Learn to separate data with simple decision boundary

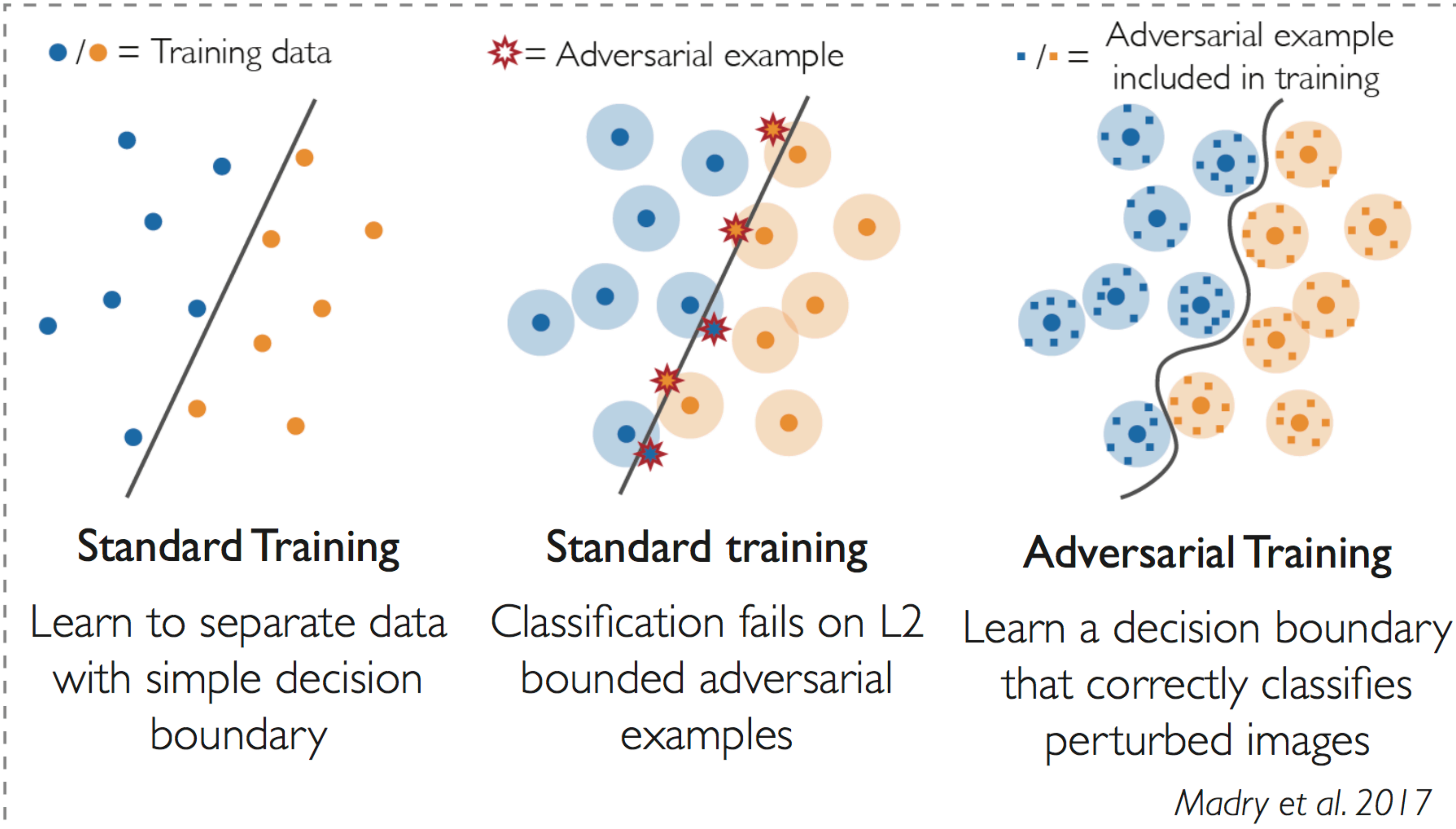
Standard training

Classification fails on L2 bounded adversarial examples

Models can be fooled by small (imperceptible to humans) adversarial perturbations.

How to address model inadequacies?

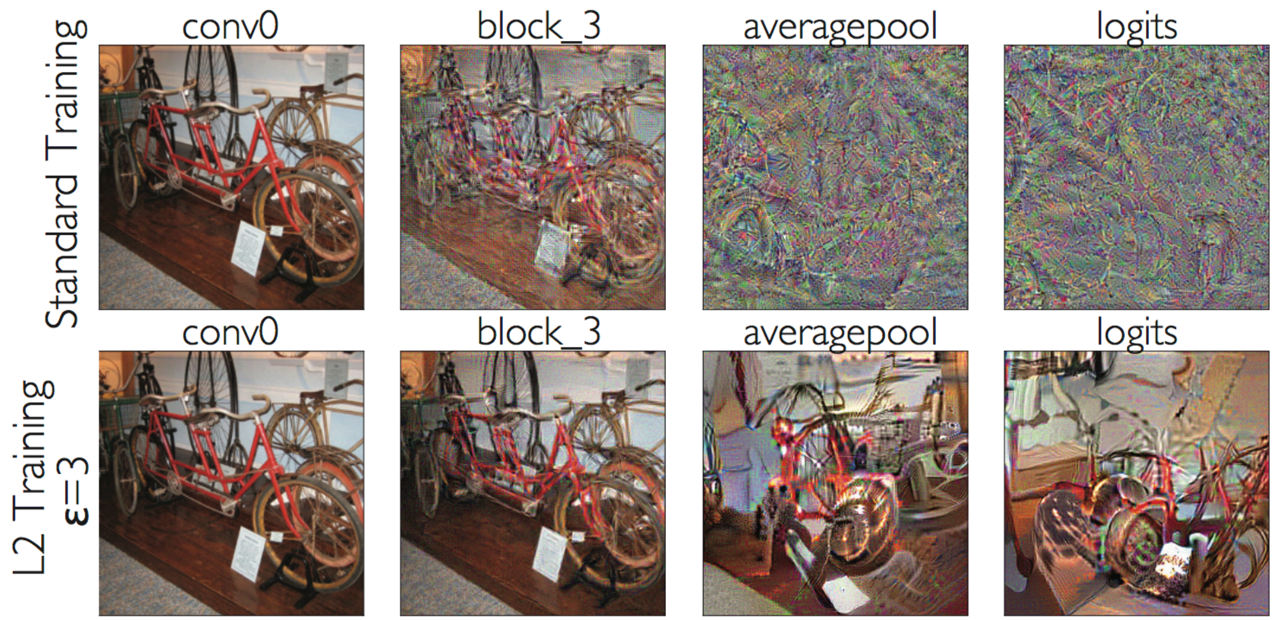
Adversarial robustness: Adversarial examples generated during training; model is trained to correctly classify them



How to address model inadequacies?

Adversarial robustness: Adversarial examples generated during training; model is trained to correctly classify them

Robust models have metamers that are more recognizable to humans:



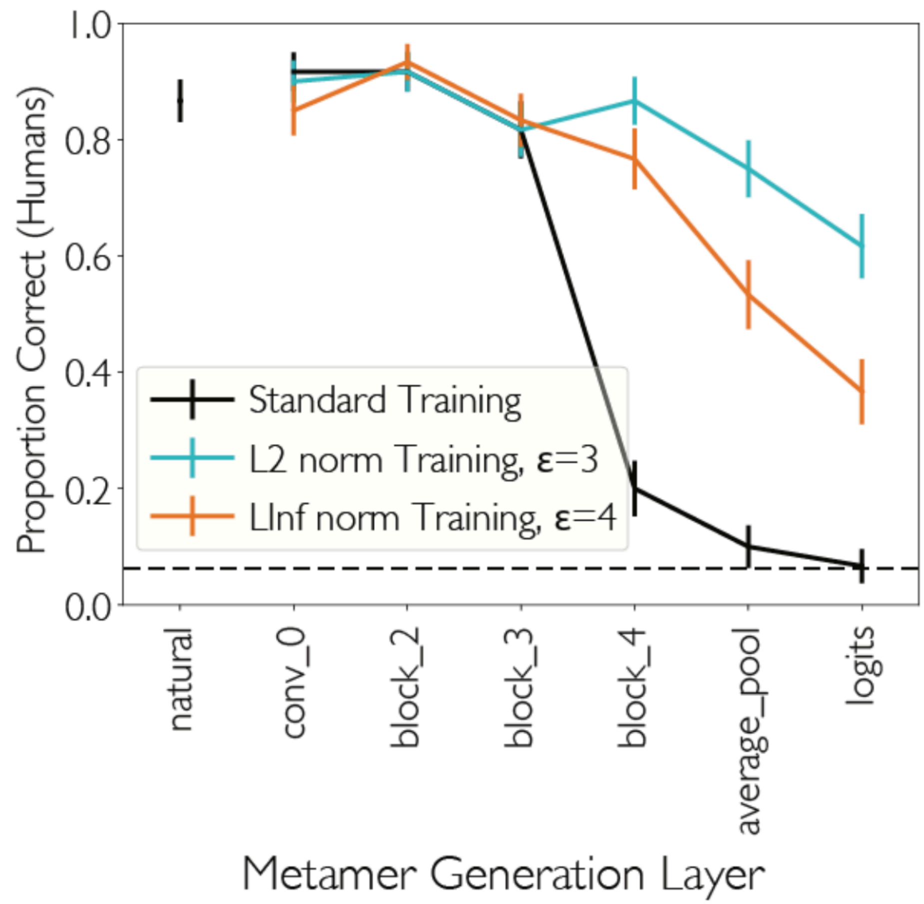
Joint work with Guillaume Leclerc, Aleksander Madry

How to address model inadequacies?

Adversarial robustness: Adversarial examples generated during training; model is trained to correctly classify them

Robust models have metamers that are more recognizable to humans:

Visual recognition (humans)



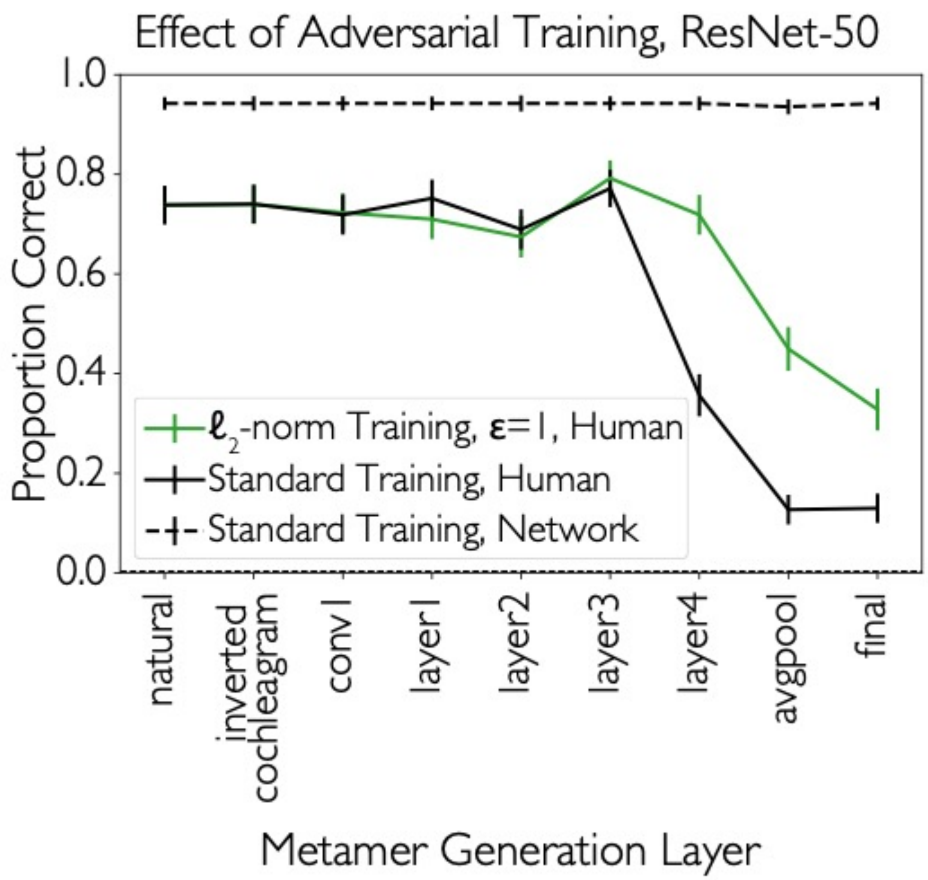
How to address model inadequacies?

Adversarial robustness: Adversarial examples generated during training; model is trained to correctly classify them

Robust models have metamers that are more recognizable to humans:

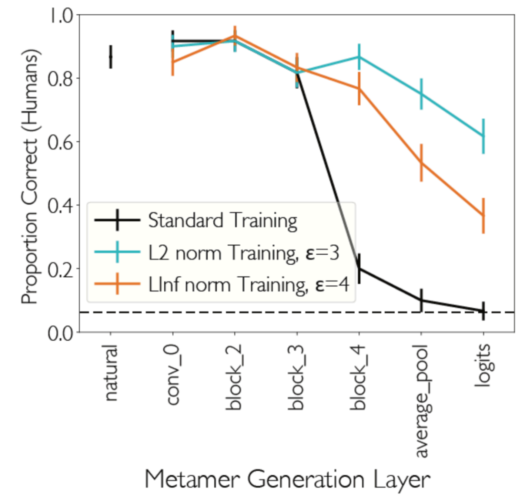
Though issue is far from completely fixed.

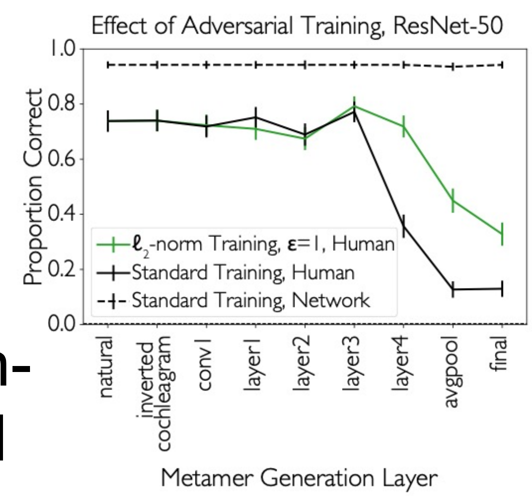
Auditory recognition (humans)



Why does adversarial training produce more human-recognizable metamers?

- Metamers are a bit like the converse of adversarial examples
  - Model judges them to be the same, but they look/sound different to humans
- But independent of a classifier
  - Just as relevant for models trained without supervision
- Not obvious why forcing invariance to human-imperceptible perturbations eliminates model invariances that humans lack…

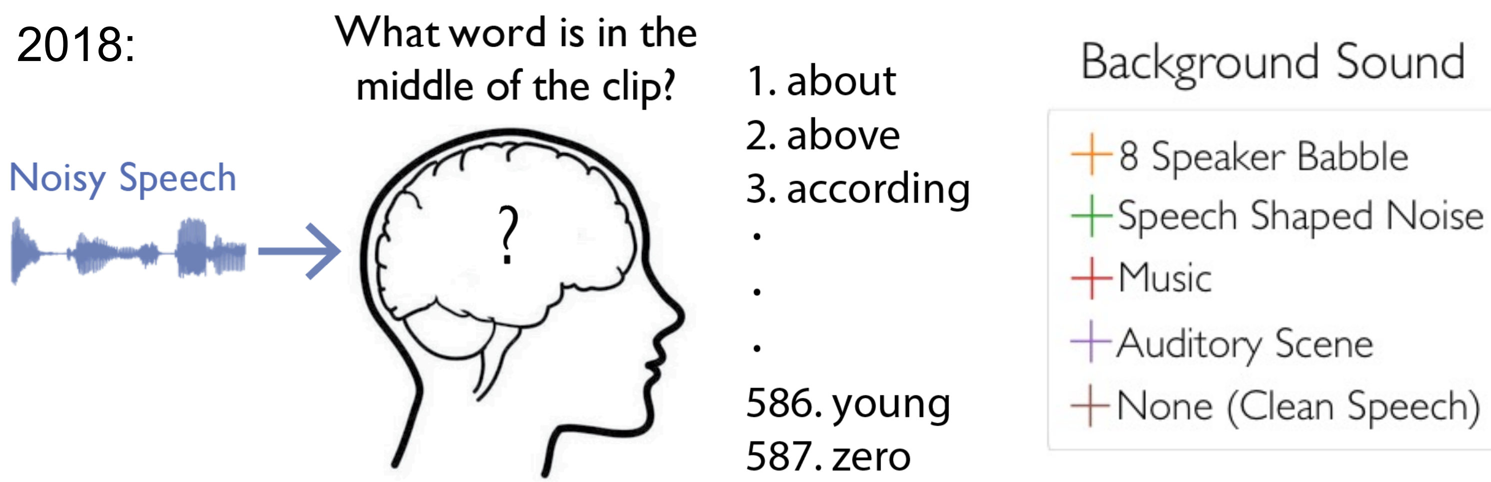




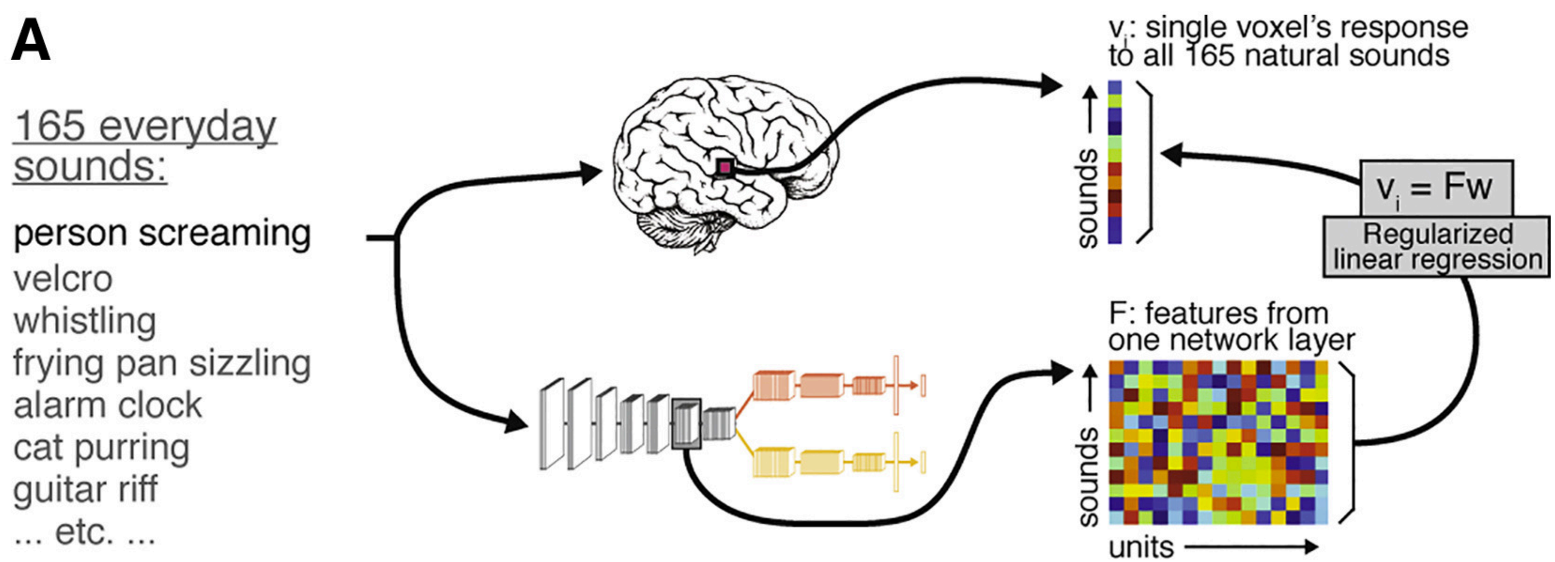
Metamers reveal differences not evident with our usual metrics

From Kell et al. 2018:

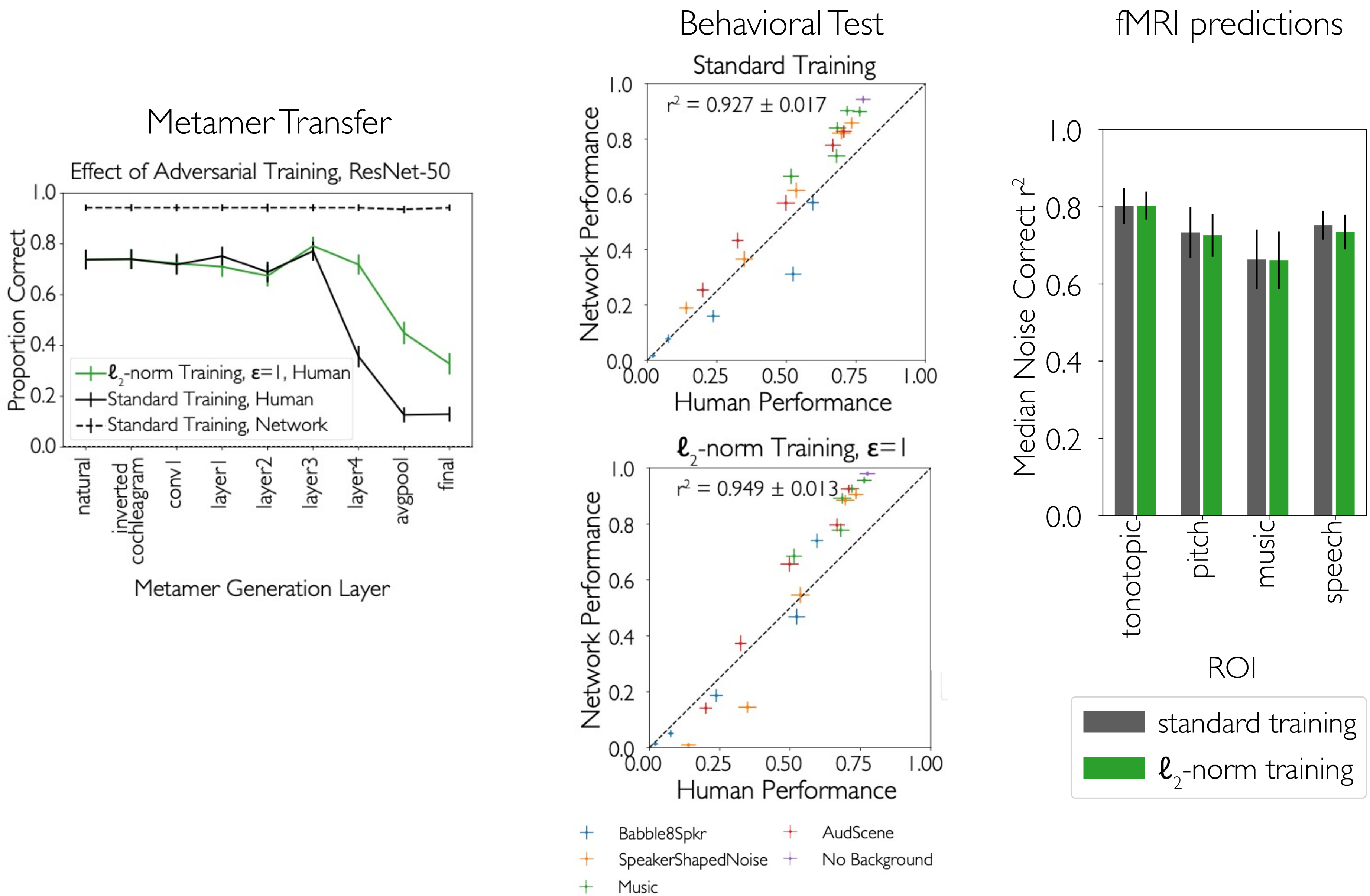
Recognition of natural sounds



fMRI predictions



Metamers reveal differences not evident with our usual metrics



Take-Home Messages, Part 2

Metamers of neural networks provide a way to reveal model invariances

- Metamers of deep layers of standard neural network models are not metameric for humans
  - Not even recognizable to humans
  - True for vision and auditory networks

- Model metamers can be made more human-recognizable with some architectural modifications (reducing aliasing)

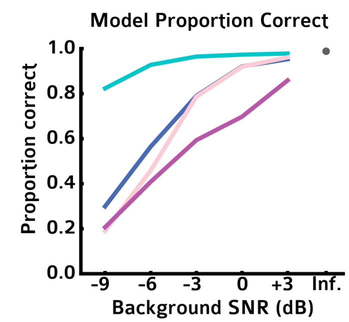
  - And by making models more robust to adversarial examples (for reasons we don’t yet fully understand)

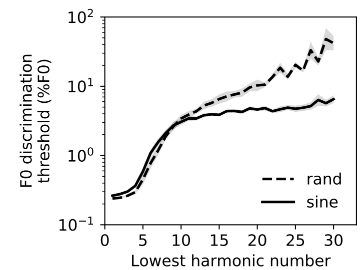
  - But divergences remain

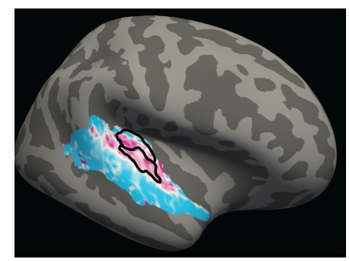
Summary

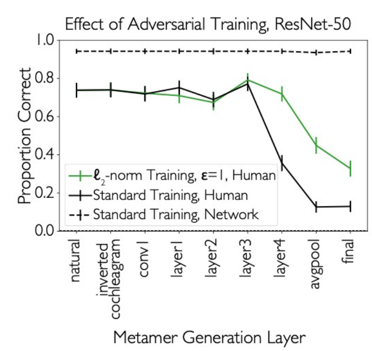
New models via deep learning of audio tasks

- Compelling matches to human behavior with real-world sounds and tasks
- And for many classical psychophysical results
- Insight into origins of behavioral traits
- Better models of auditory cortex
- Evidence for hierarchical organization
- Significant remaining discrepancies revealed with model metamers









ACKNOWLEDGMENTS

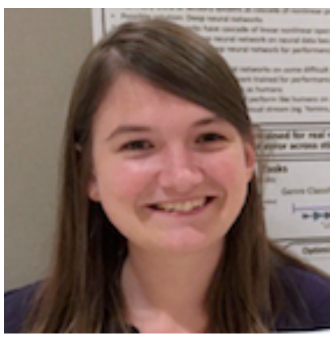
Alex Kell

Andrew Francl

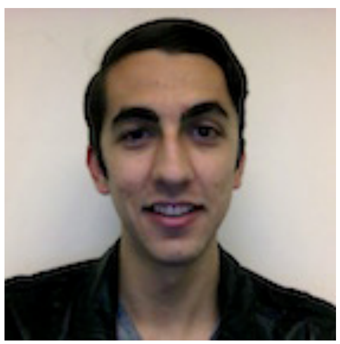
Mark Saddler

Jenelle Feather

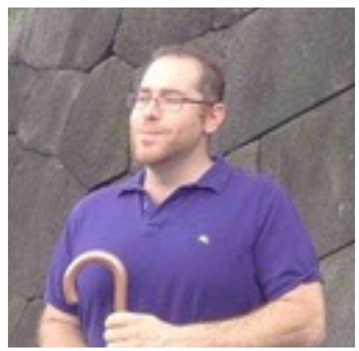
Erica Shook



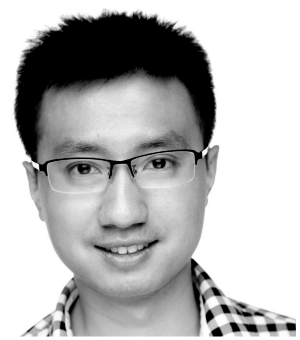
Ray Gonzalez



Dan Yamins



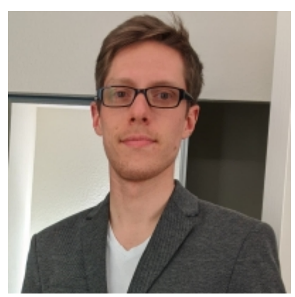
Yang Zhang



Kaizhi Qian

National Science Foundation
NIDCD
IBM

Guillaume Leclerc



Aleksander Madry