Neural Representations of the Cocktail Party in Human Auditory Cortex

Jonathan Z. Simon

Department of Biology
Department of Electrical & Computer Engineering
Institute for Systems Research

University of Maryland

Acknowledgements

Grad Students
Francisco Cervantes
Alex Presacco
Krishna Puvvada

Past Grad Students
Nayef Ahmar
Claudia Bonin
Maria Chait
Marisel Villafane Delgado
Kim Drnec
Nai Ding
Victor Grau-Serrat
Ling Ma
Raul Rodriguez
Juanjuan Xiang
Kai Sum Li
Jiachen Zhuo

Undergraduate Students
Abdulaziz Al-Turki
Nicholas Asendorf
Sonja Bohr
Elizabeth Camenga
Corinne Cameron
Julien Dagenais
Katya Dombrowski
Kevin Hogan
Kevin Kahn
Andrea Shome
Madeleine Varmer
Ben Walsh

Collaborators' Students
Murat Aytekin
Julian Jenkins
David Klein
Huan Luo

Past Postdocs
Dan Hertz
Yadong Wang

Collaborators
Catherine Carr
Monita Chatterjee
Alain de Cheveigné
Didier Depireux
Mounya Elhilali
Jonathan Fritz
Cindy Moss
David Poeppel
Shihab Shamma

The Cocktail Party

Alex Katz,
The Cocktail Party

Introduction

- Magnetoencephalography (MEG)

- Cortical Representations of Speech

  - Decoding vs. Encoding

  - Attended vs. Unattended Speech

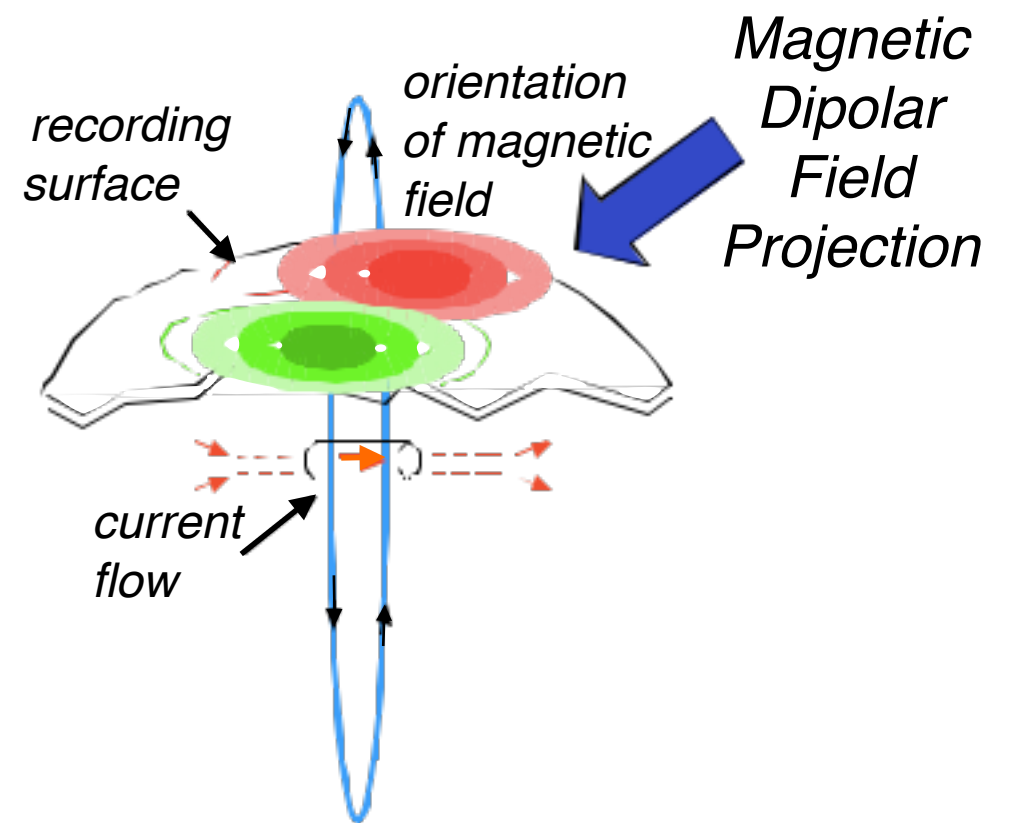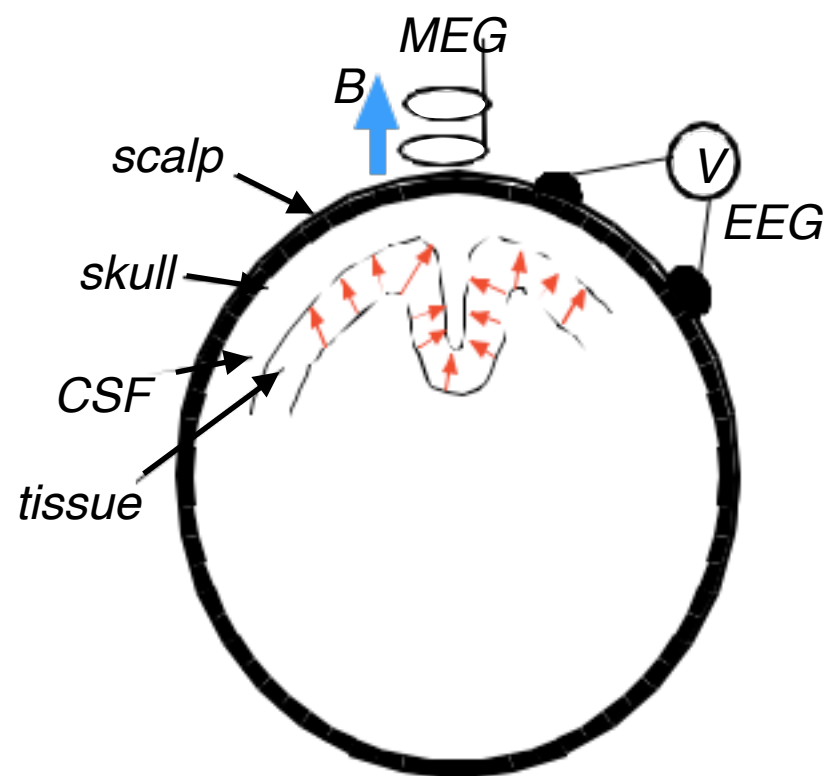 - Foreground vs. Background

Magnetoencephalography

- Non-invasive, Passive, Silent Neural Recordings

- Simultaneous Whole-Head Recording (~200 sensors)

- Sensitivity
  - high: ~100 fT (10^{-13} Tesla)
  - low: ~10^4 – ~10^6 neurons

- Temporal Resolution: ~1 ms

- Spatial Resolution
  - coarse: ~1 cm
  - ambiguous
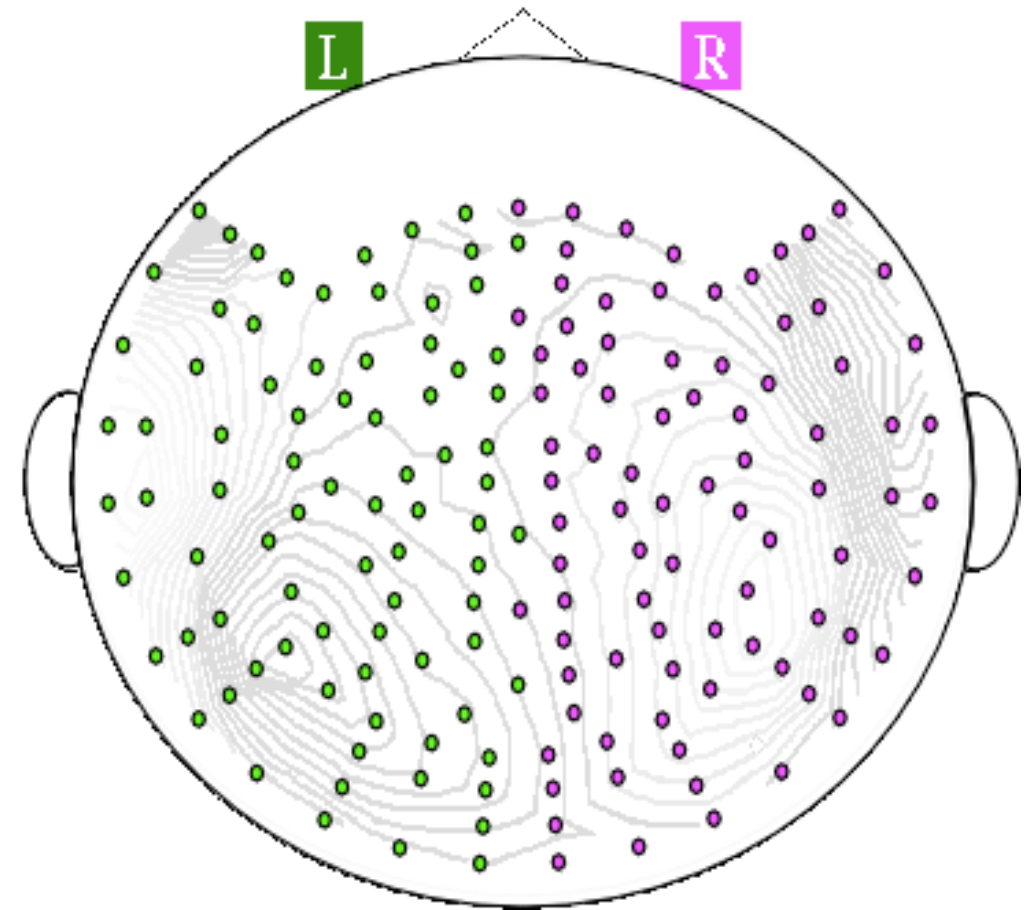
Neural Signals & MEG

Photo by Fritz Goro
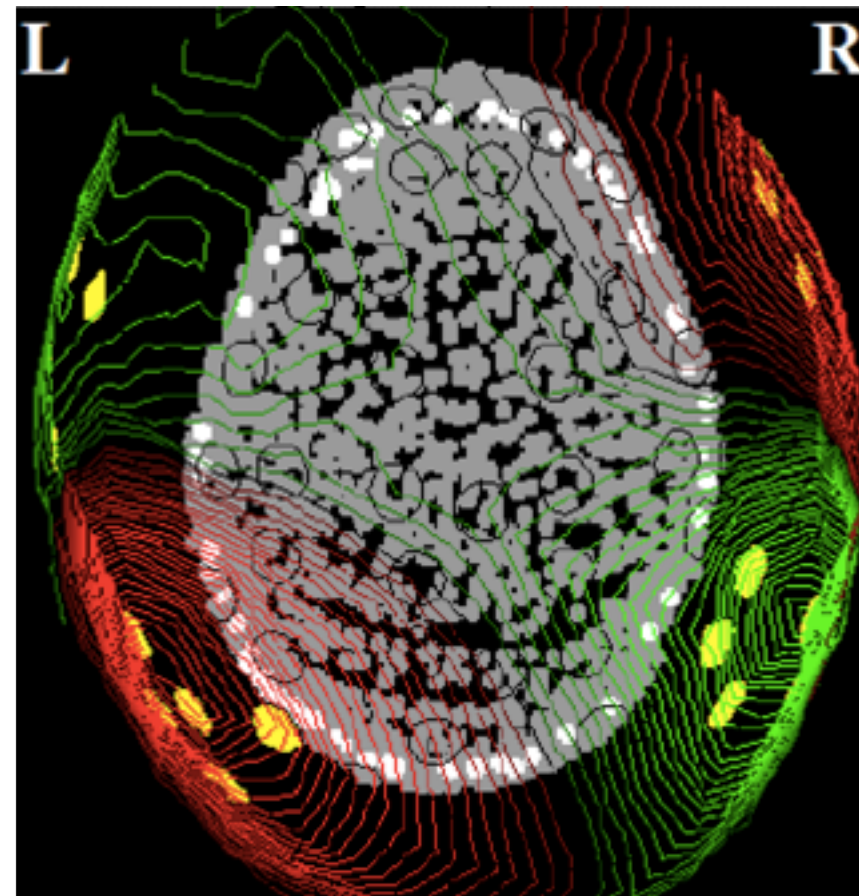
- Direct electrophysiological measurement
  - not hemodynamic
  - real-time
- No unique solution for distributed source

- Measures spatially synchronized cortical activity
- Fine temporal resolution (~ 1 ms)
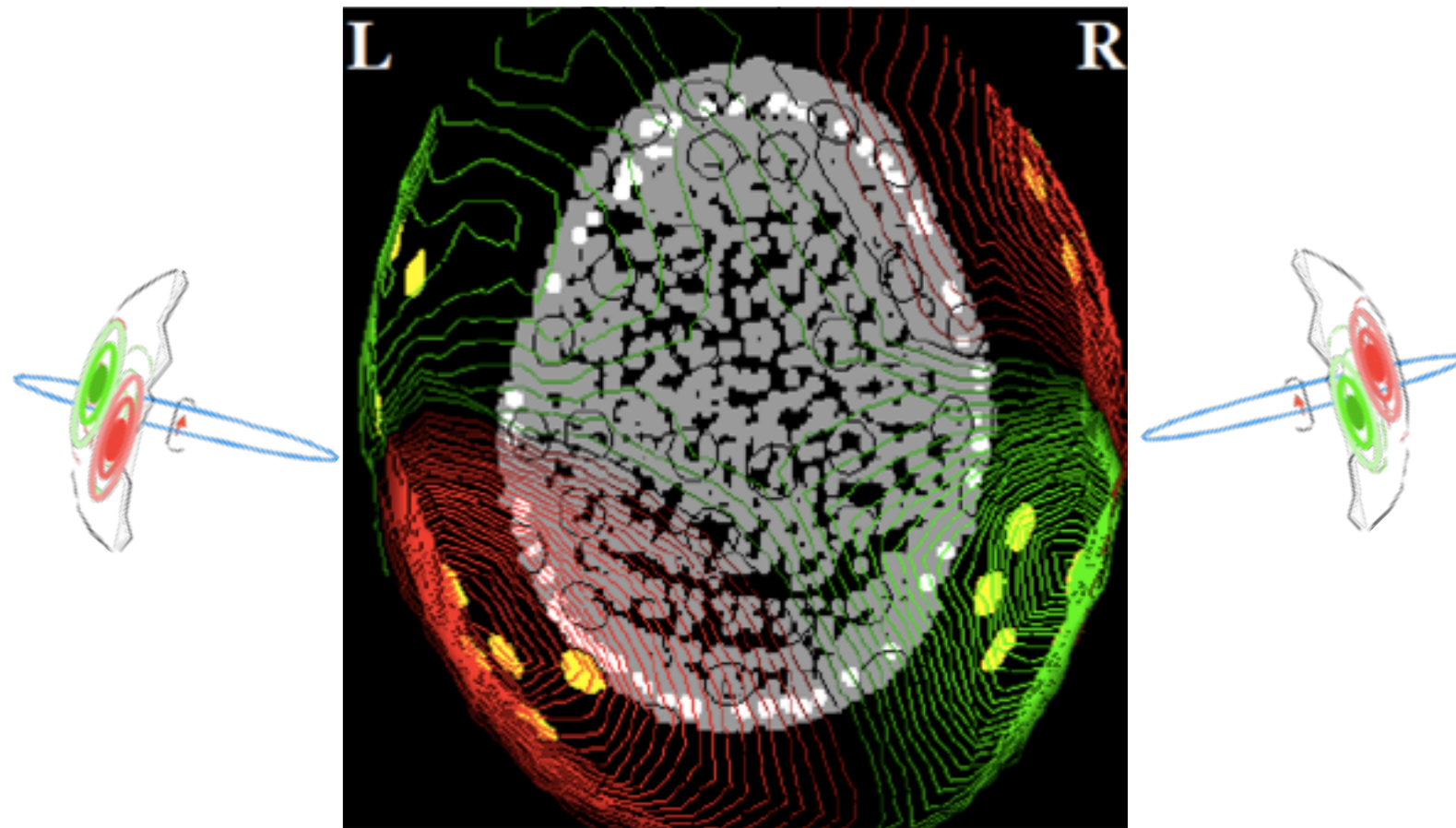- Moderate spatial resolution (~ 1 cm)

MEG Auditory Field
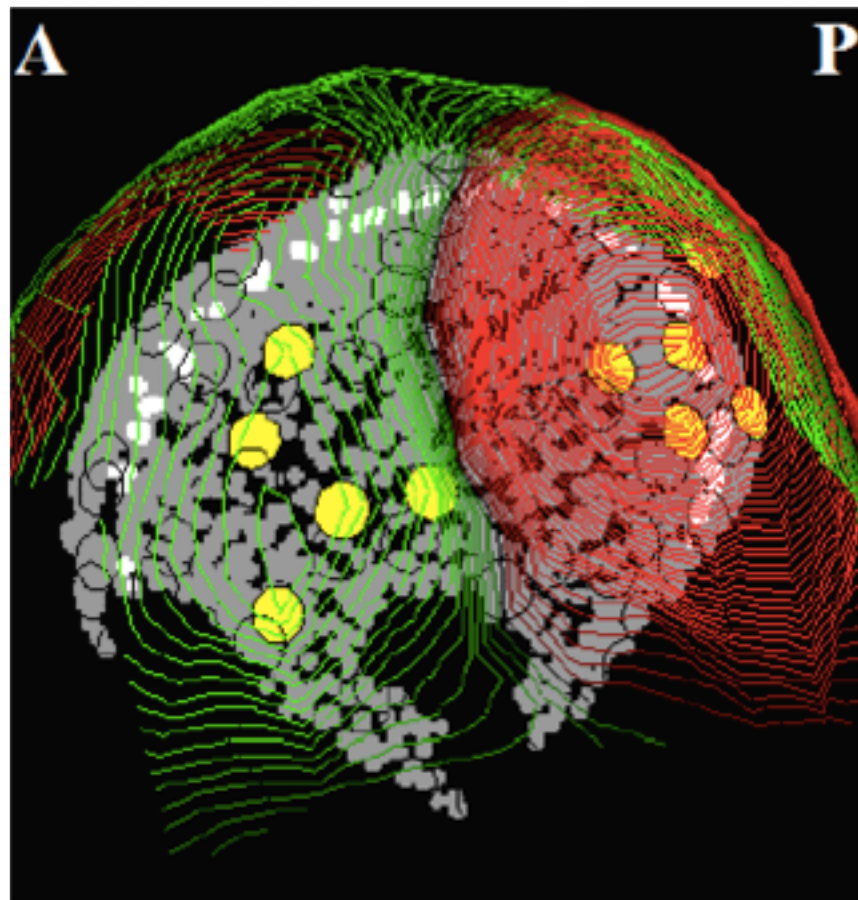
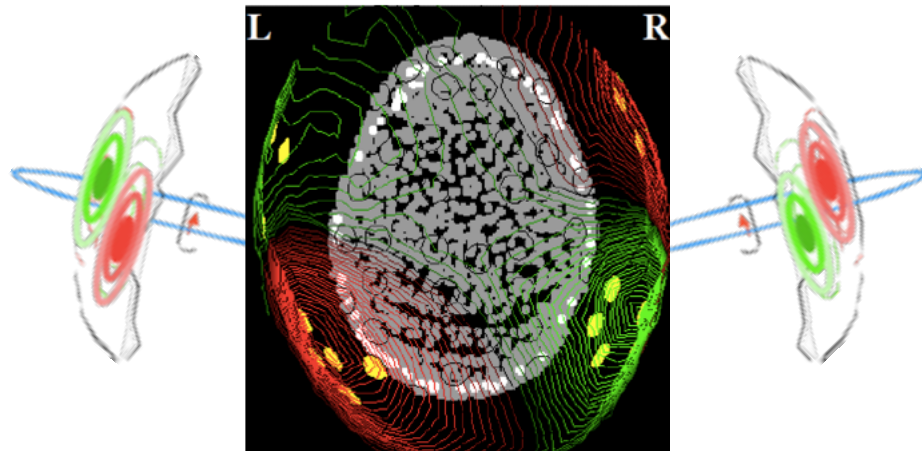
Sagittal View

Axial View

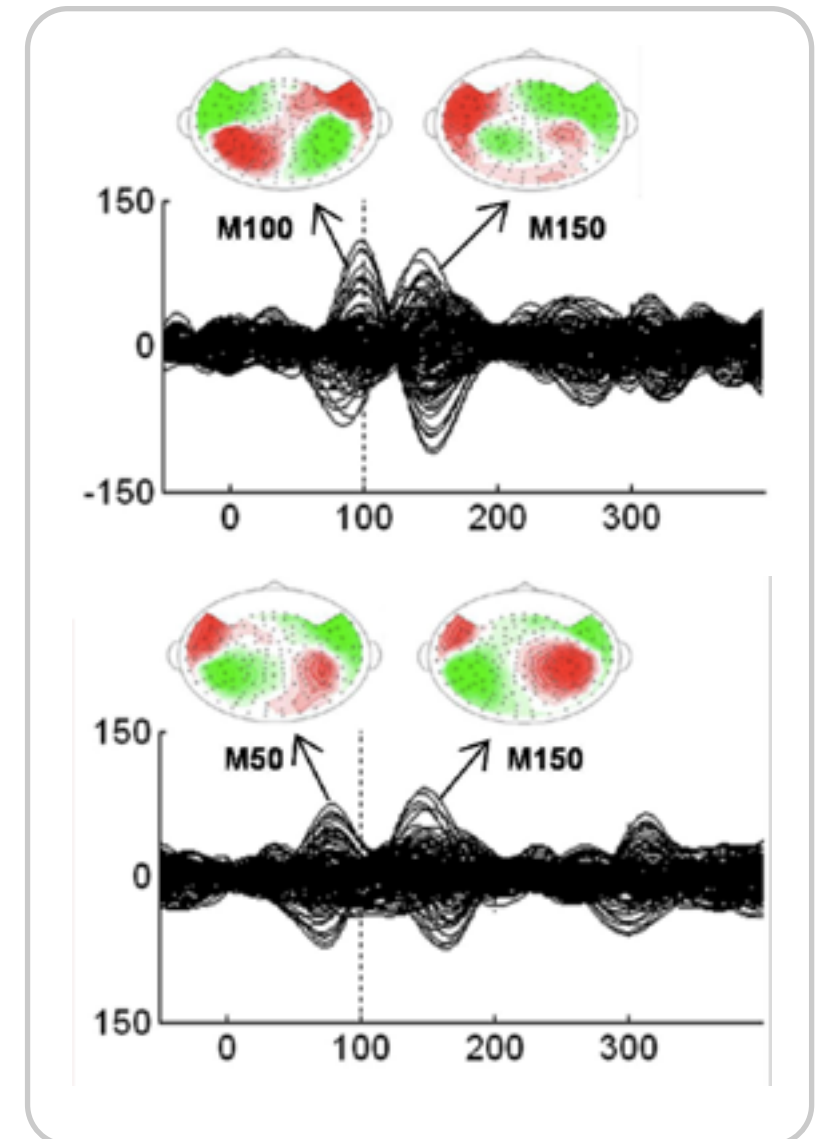Strongly
Lateralized

MEG Auditory Field

Chait et al., Cerebral Cortex (2006)

Time Course of MEG Responses

Auditory Evoked Responses

- MEG Response Patterns Time-Locked to Stimulus Events
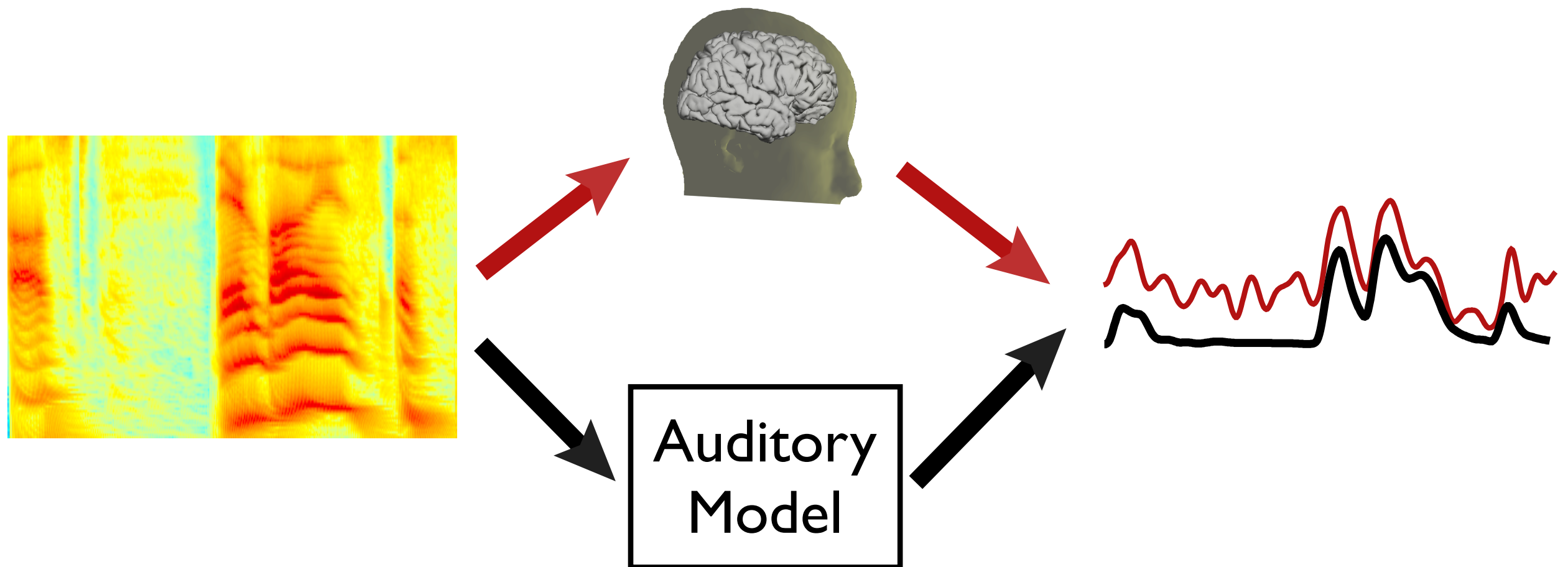
- Robust
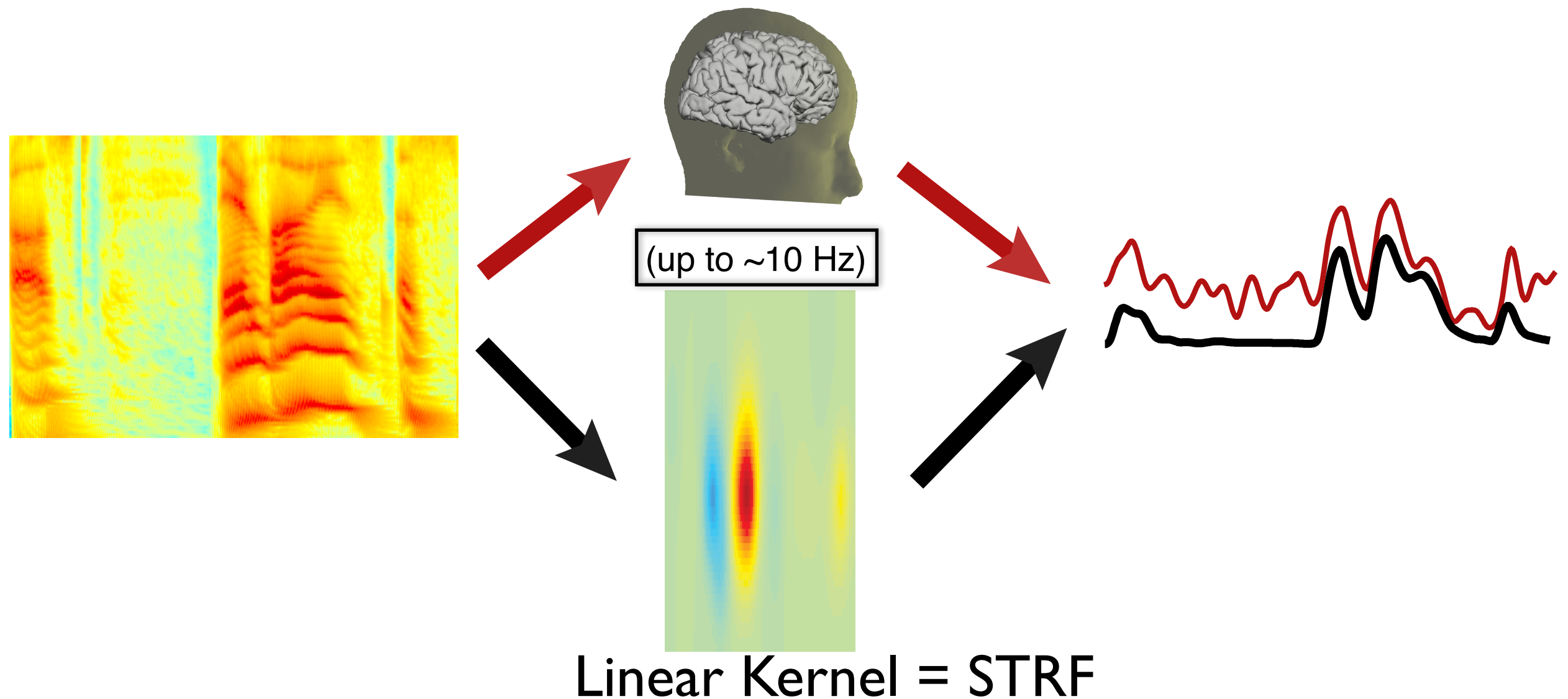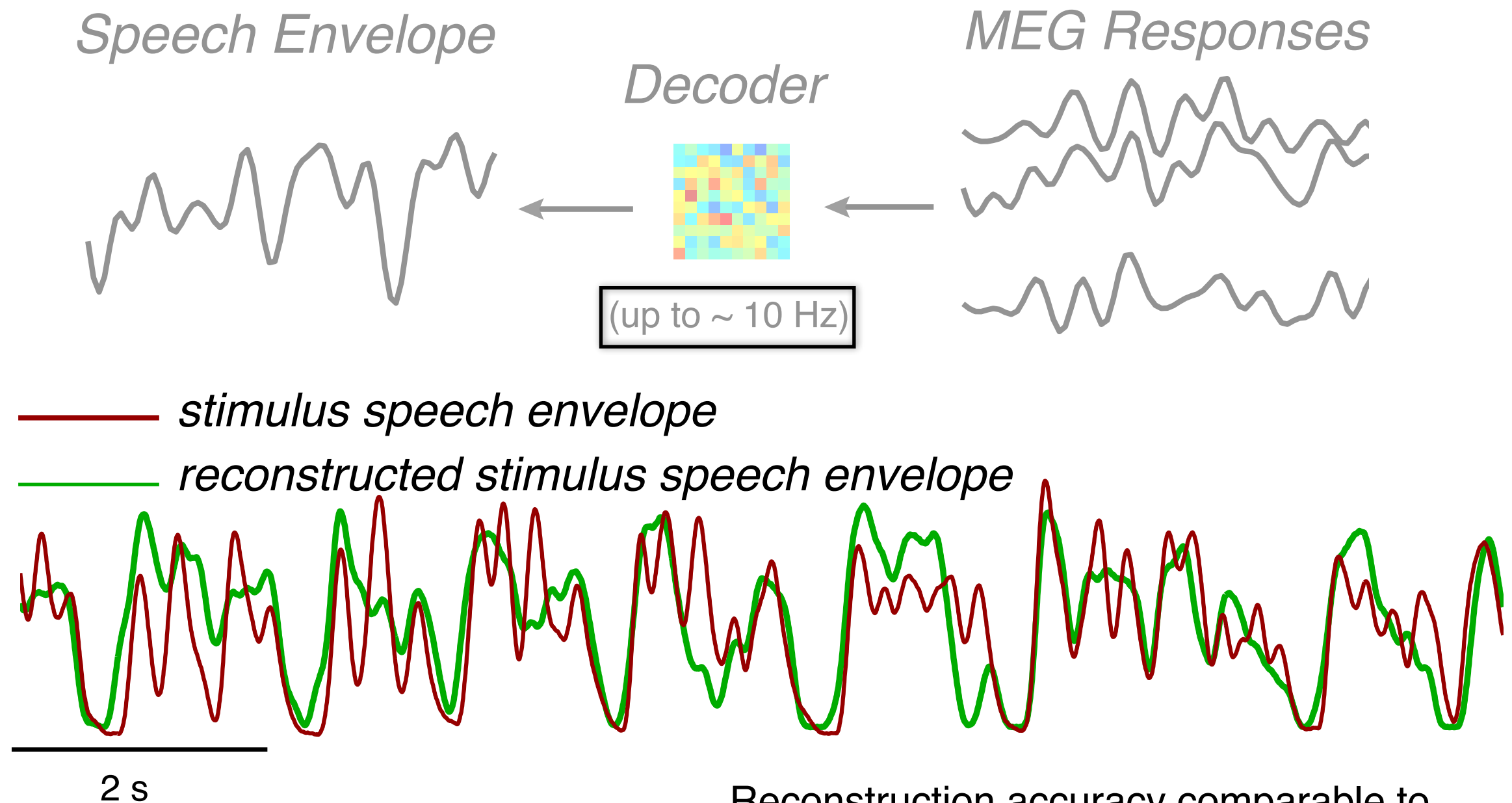
- Strongly Lateralized

Pure Tone

Broadband Noise

MEG Responses
to Speech Modulations
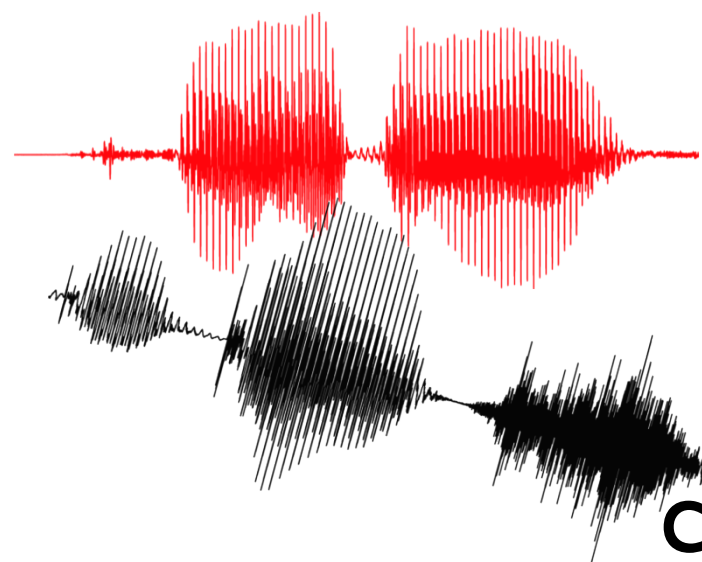
MEG Responses Predicted by STRF Model

(up to ~10 Hz)

Linear Kernel = STRF

"Spectro-Temporal Response Function"

Neural Reconstruction of Speech Envelope

Speech Envelope

Decoder

MEG Responses

(up to ~ 10 Hz)

stimulus speech envelope

reconstructed stimulus speech envelope

2 s

Reconstruction accuracy comparable to single unit & ECoG recordings

Ding & Simon, J Neurophysiol (2012)
Zion-Golumbic et al., Neuron (2013)

Experiments

Experiments in Progress

reverberation
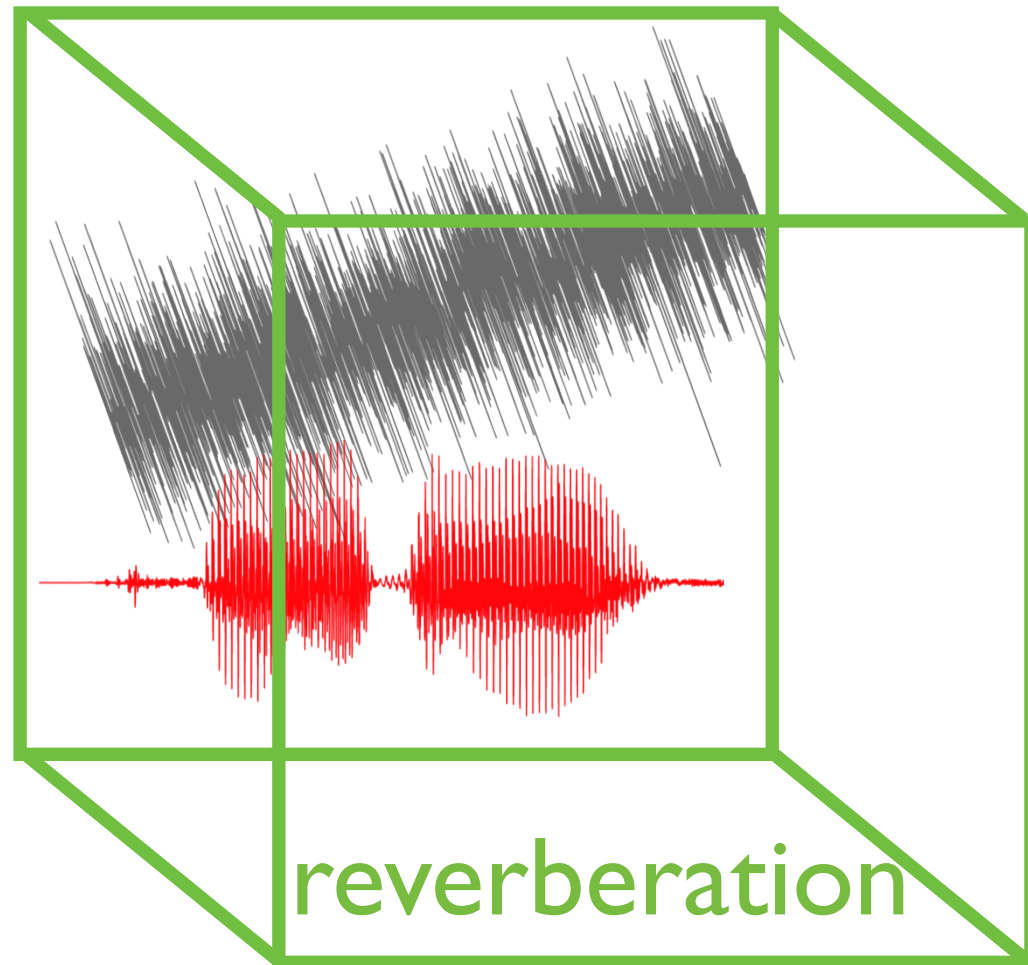
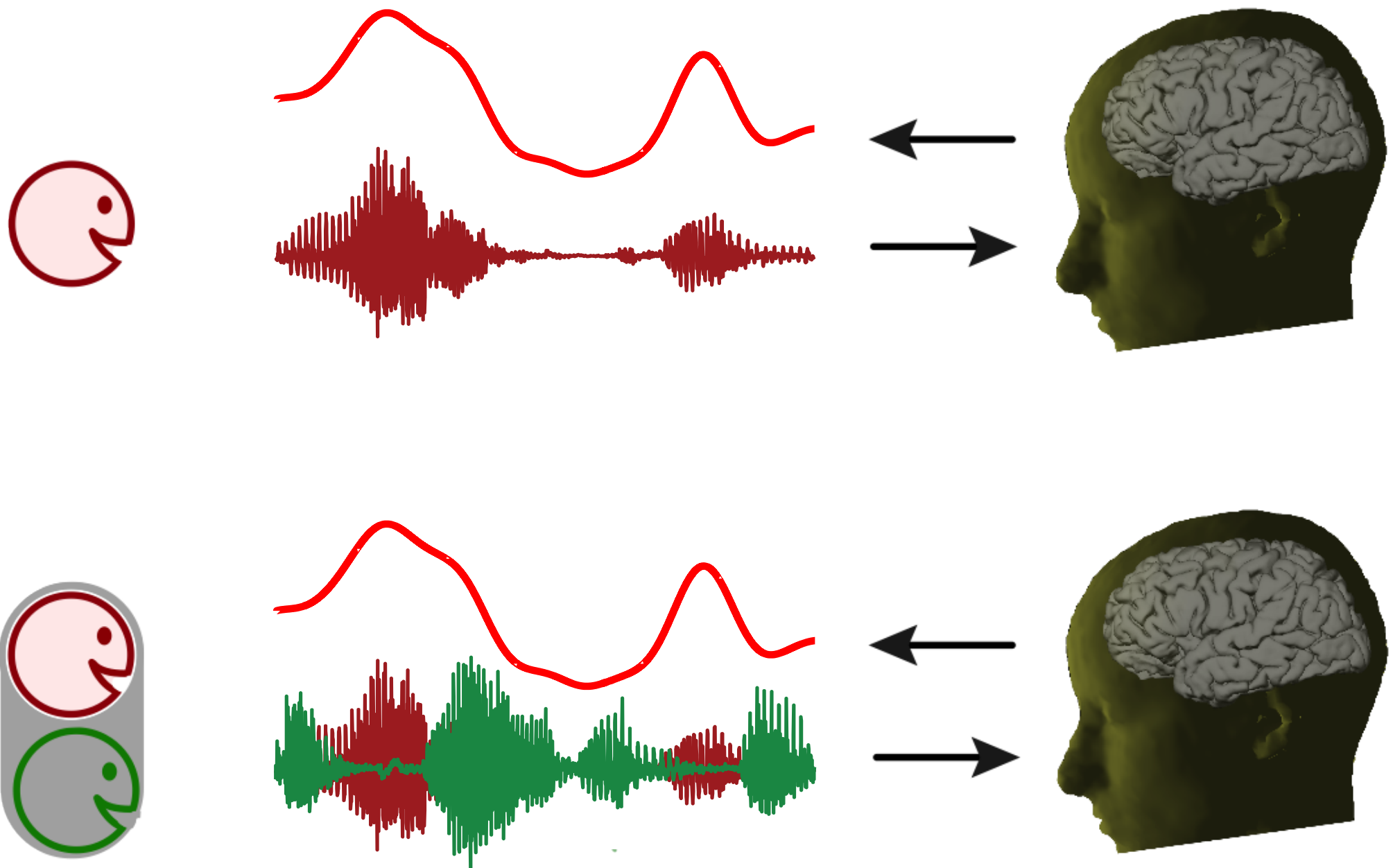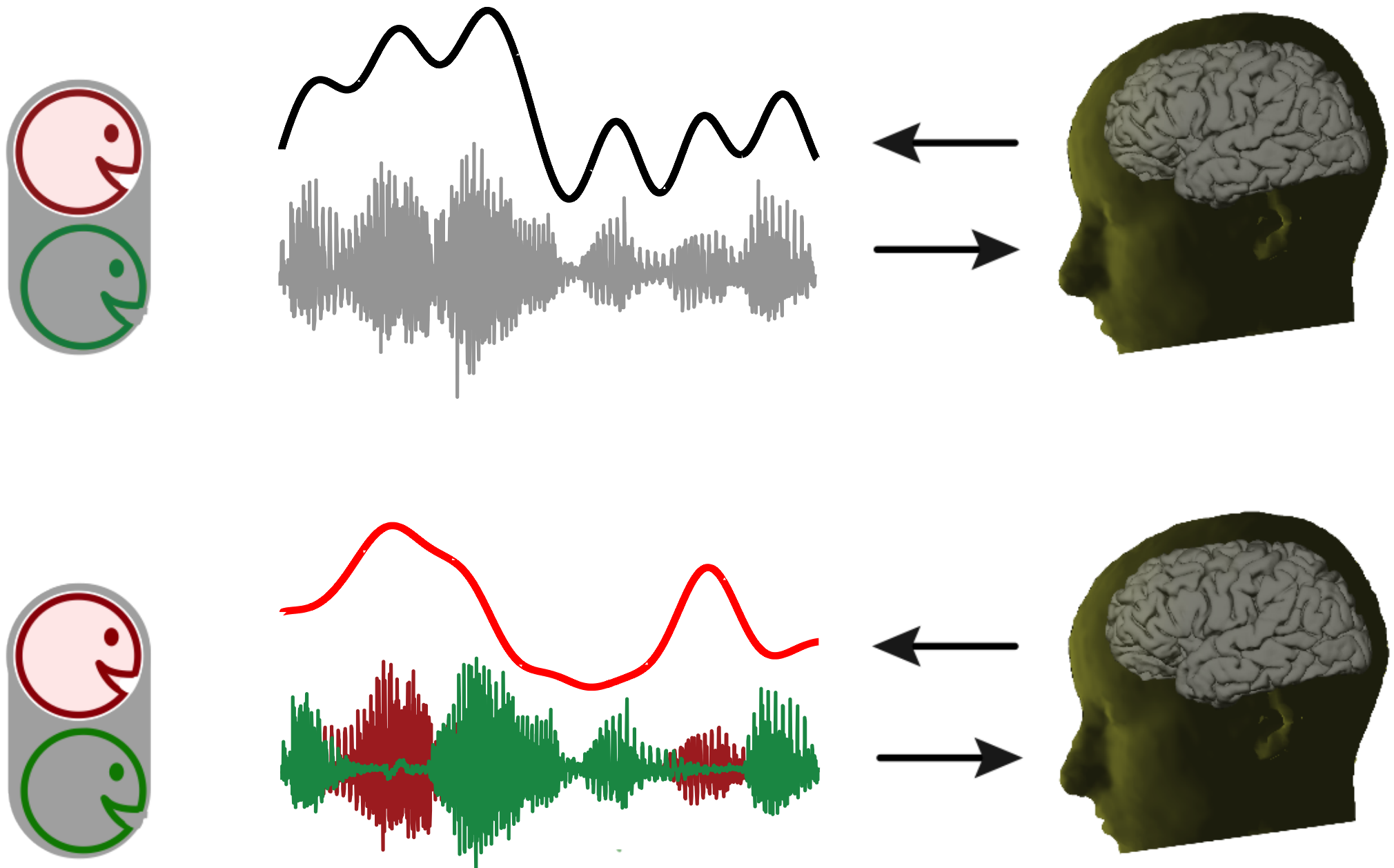# Experiments in Progress
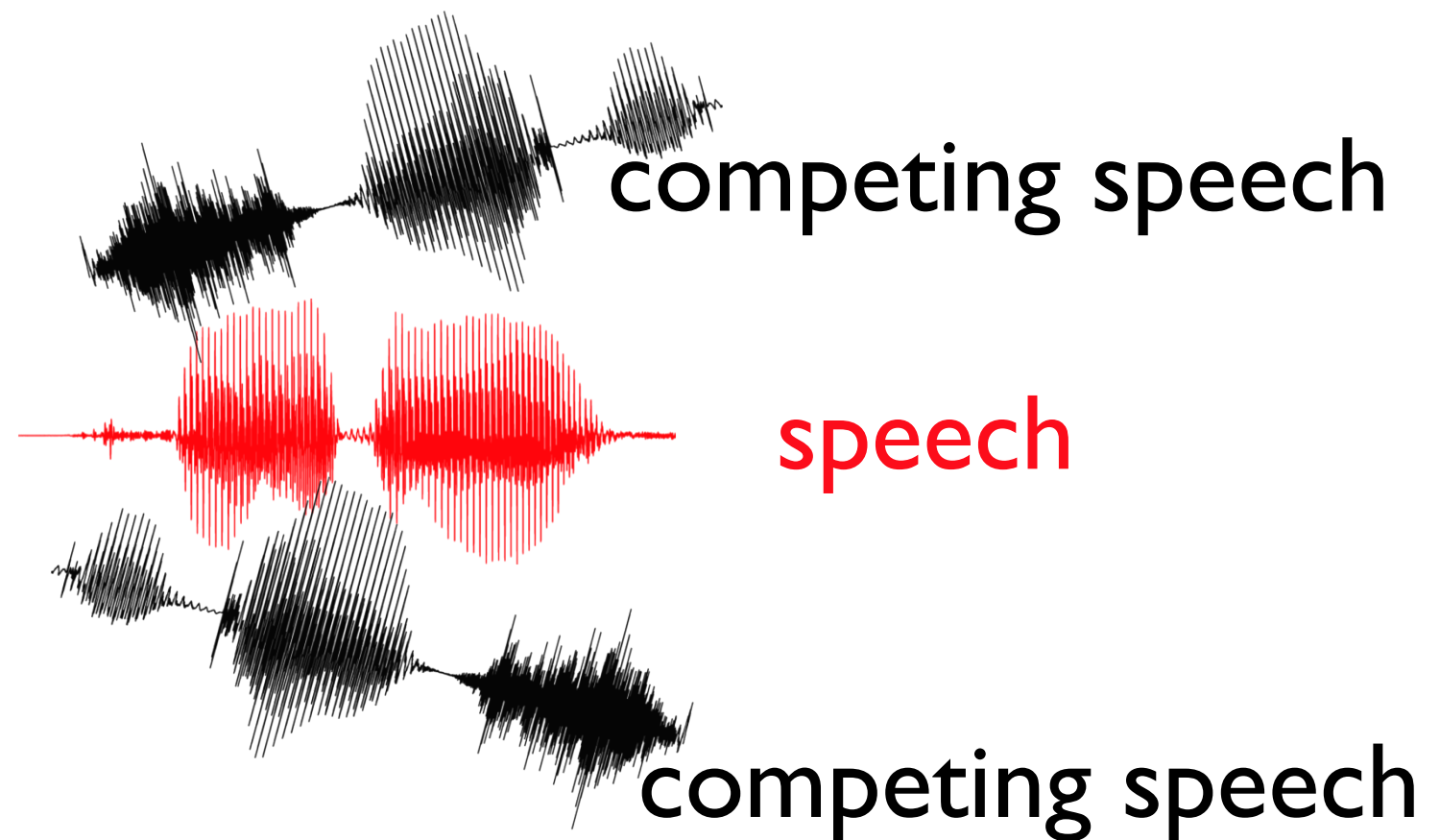
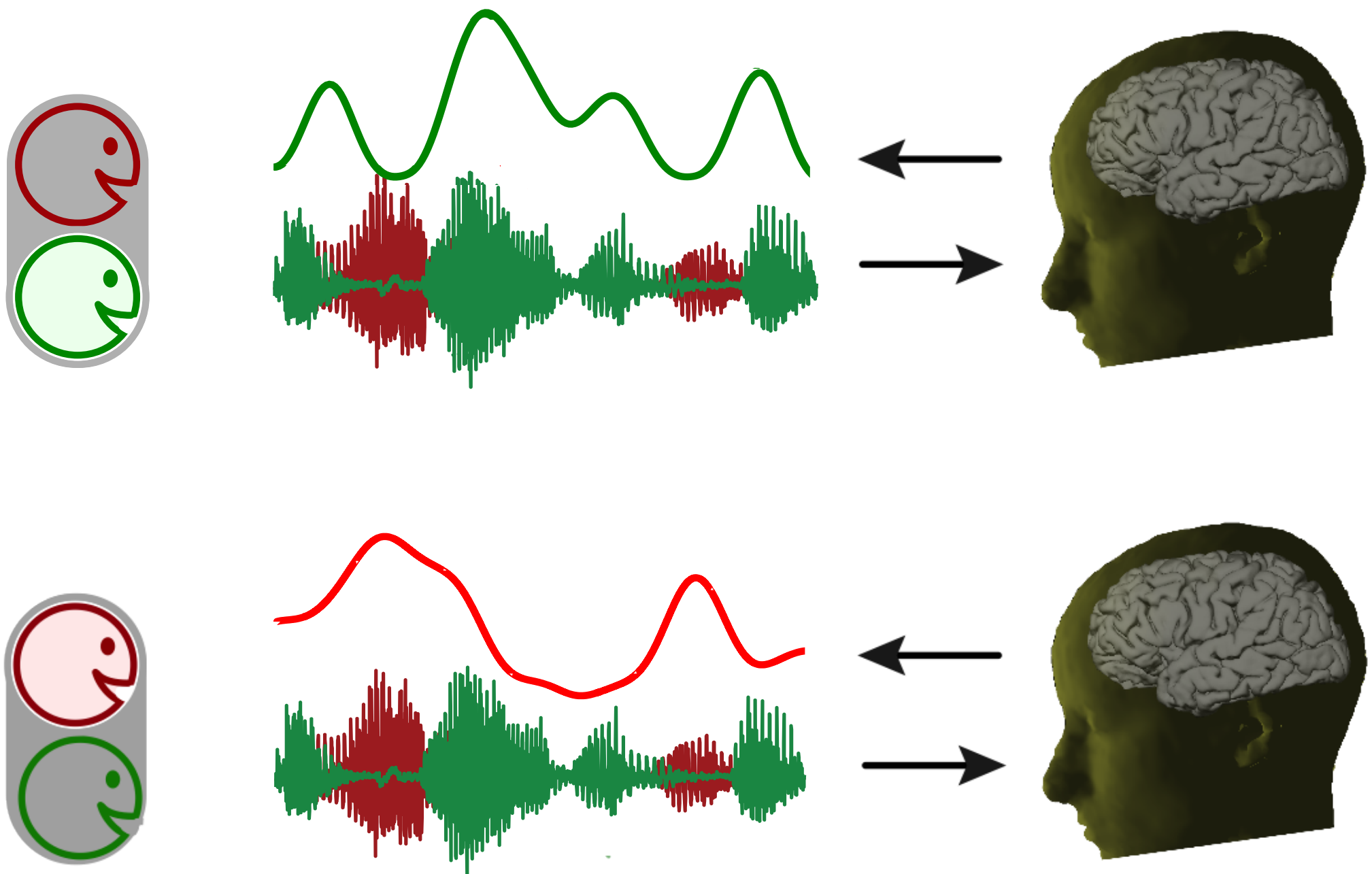# Two Competing Speakers
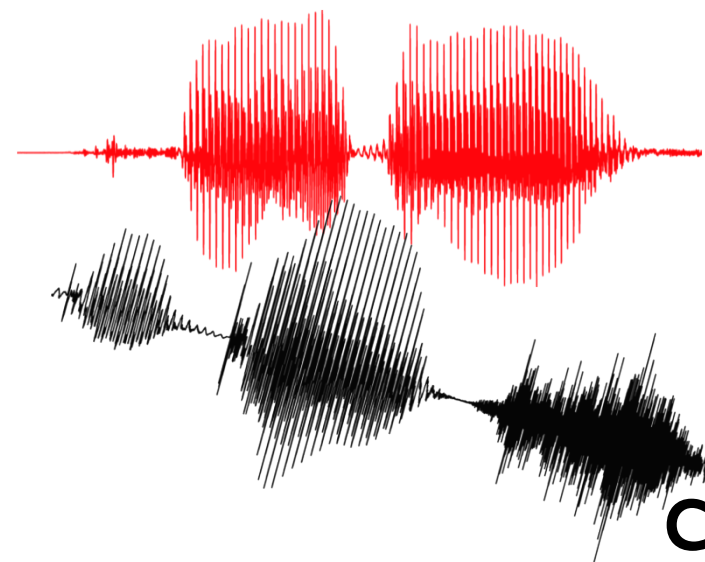
speech

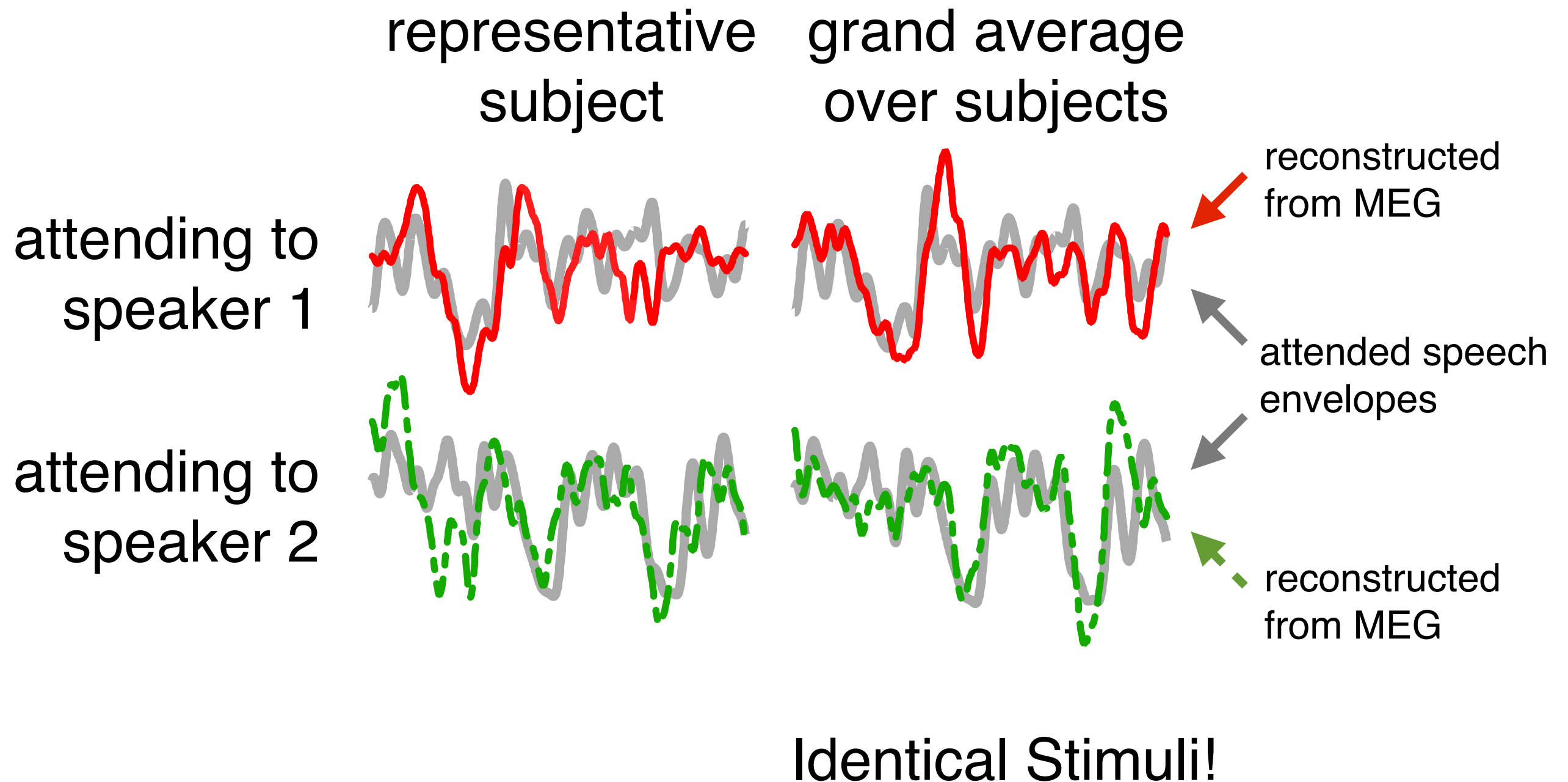competing speech

Selective Neural Encoding

Unselective vs. Selective Neural Encoding

Selective Neural Encoding

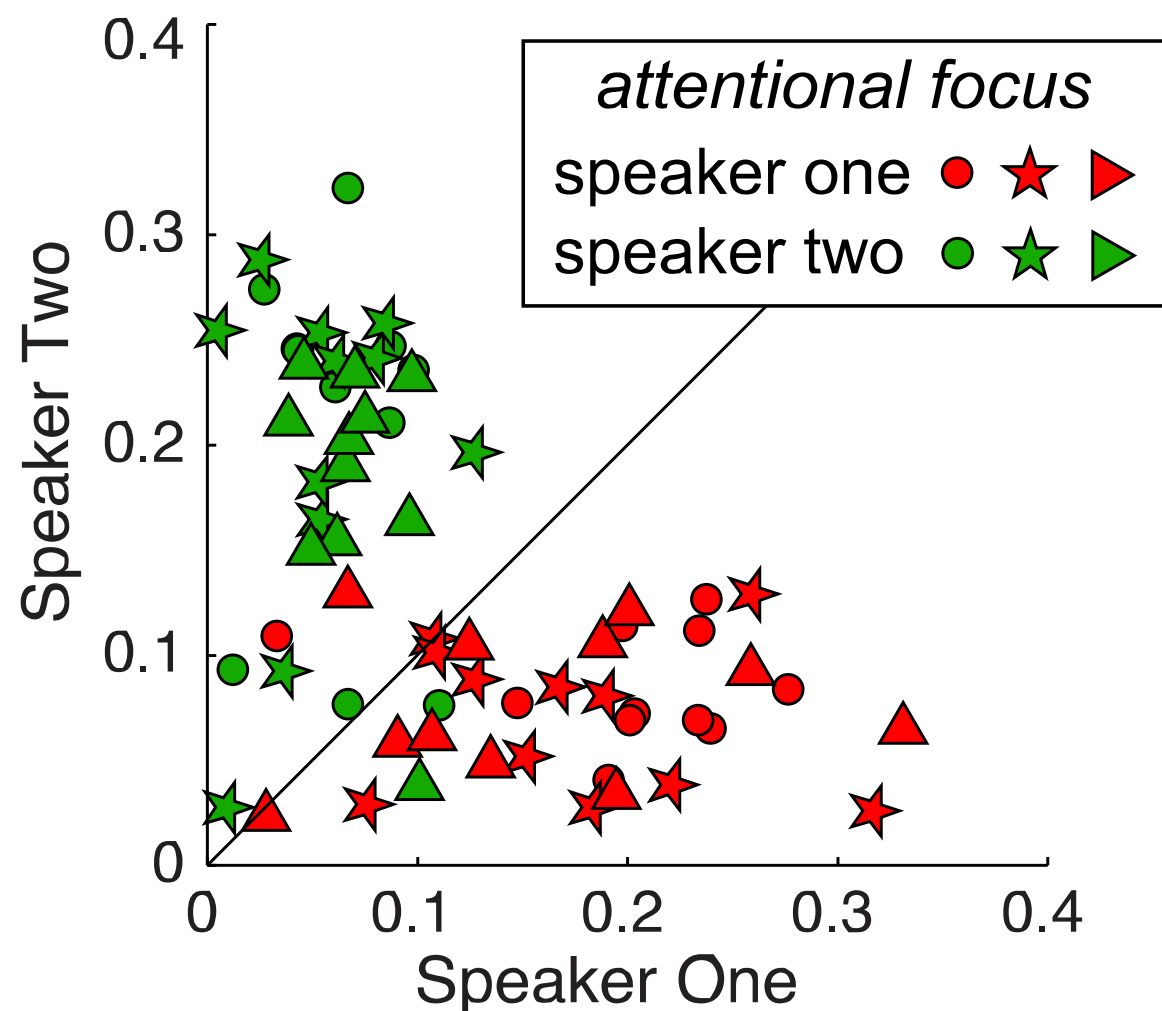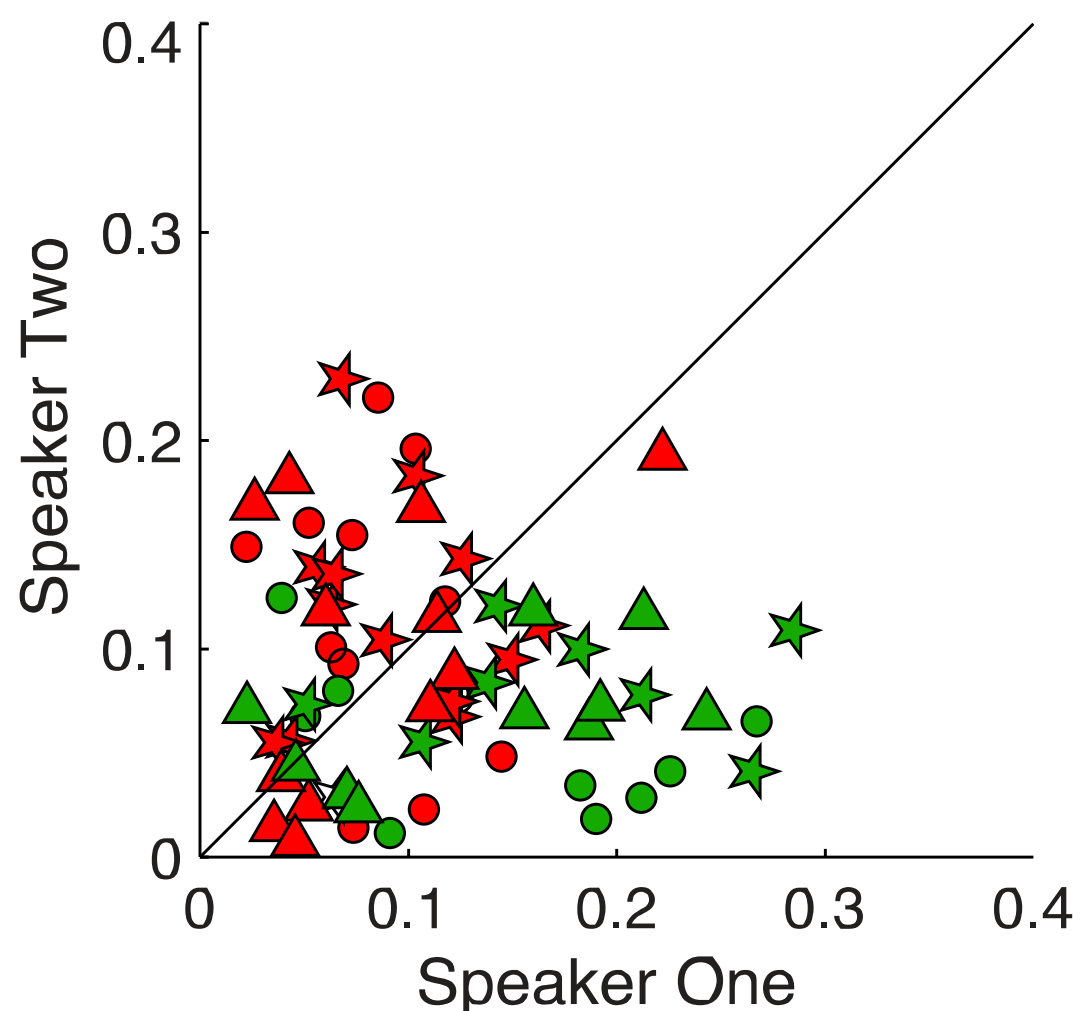Stream-Specific Representation

Ding & Simon, PNAS (2012)

Single Trial Speech Reconstruction

Ding & Simon, PNAS (2013)

Forward STRF Model

Spectro-Temporal
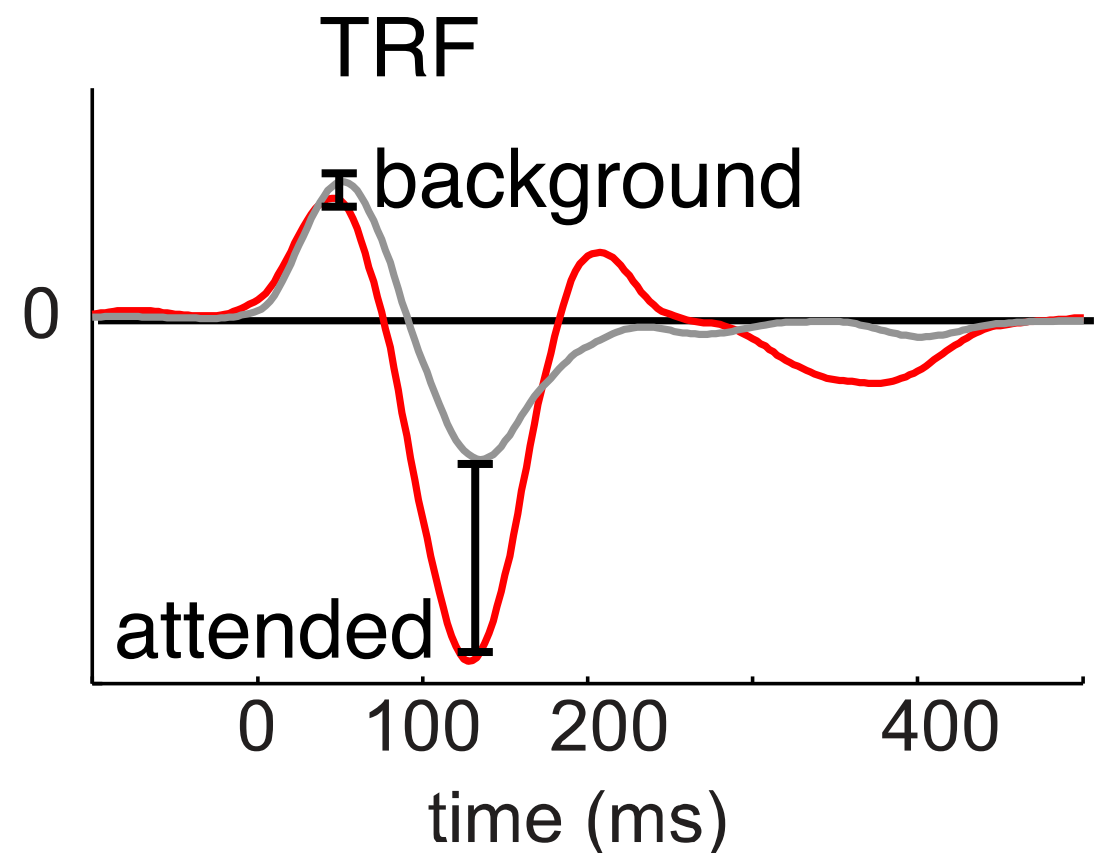Response Function
(STRF)
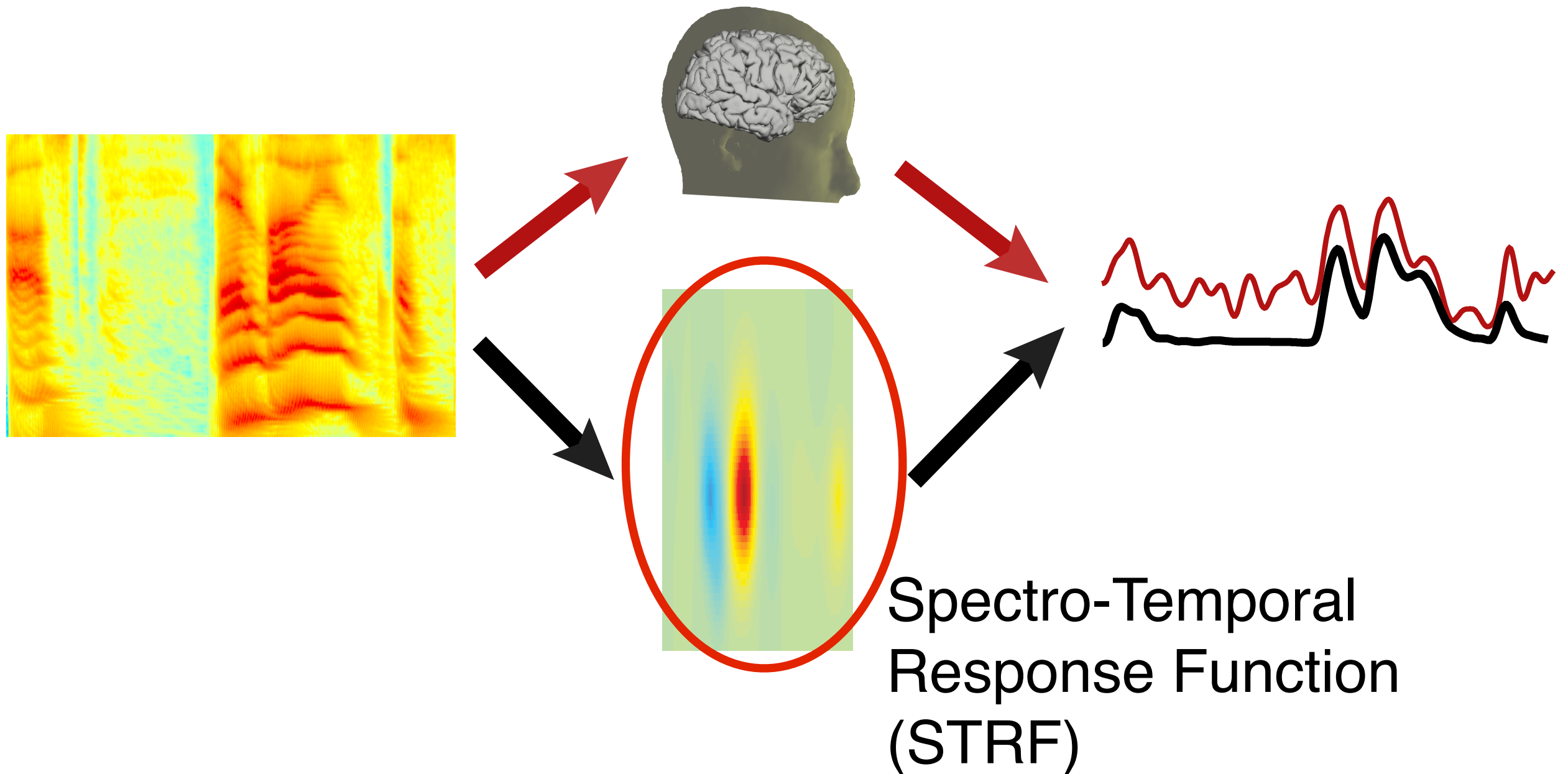
STRF Results
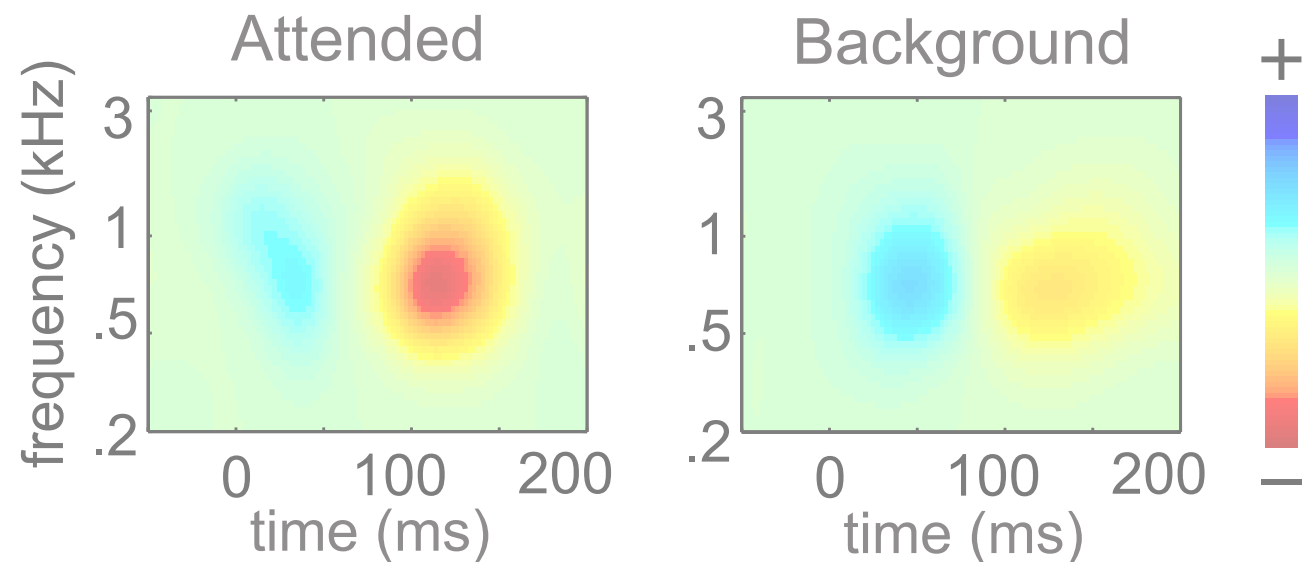
- STRF separable (time, frequency)
- 300 Hz - 2 kHz dominant carriers
- M50$_{STRF}$ positive peak
- M100$_{STRF}$ negative peak
- **M100$_{STRF}$ strongly modulated by attention,** *but not M50$_{STRF}$*
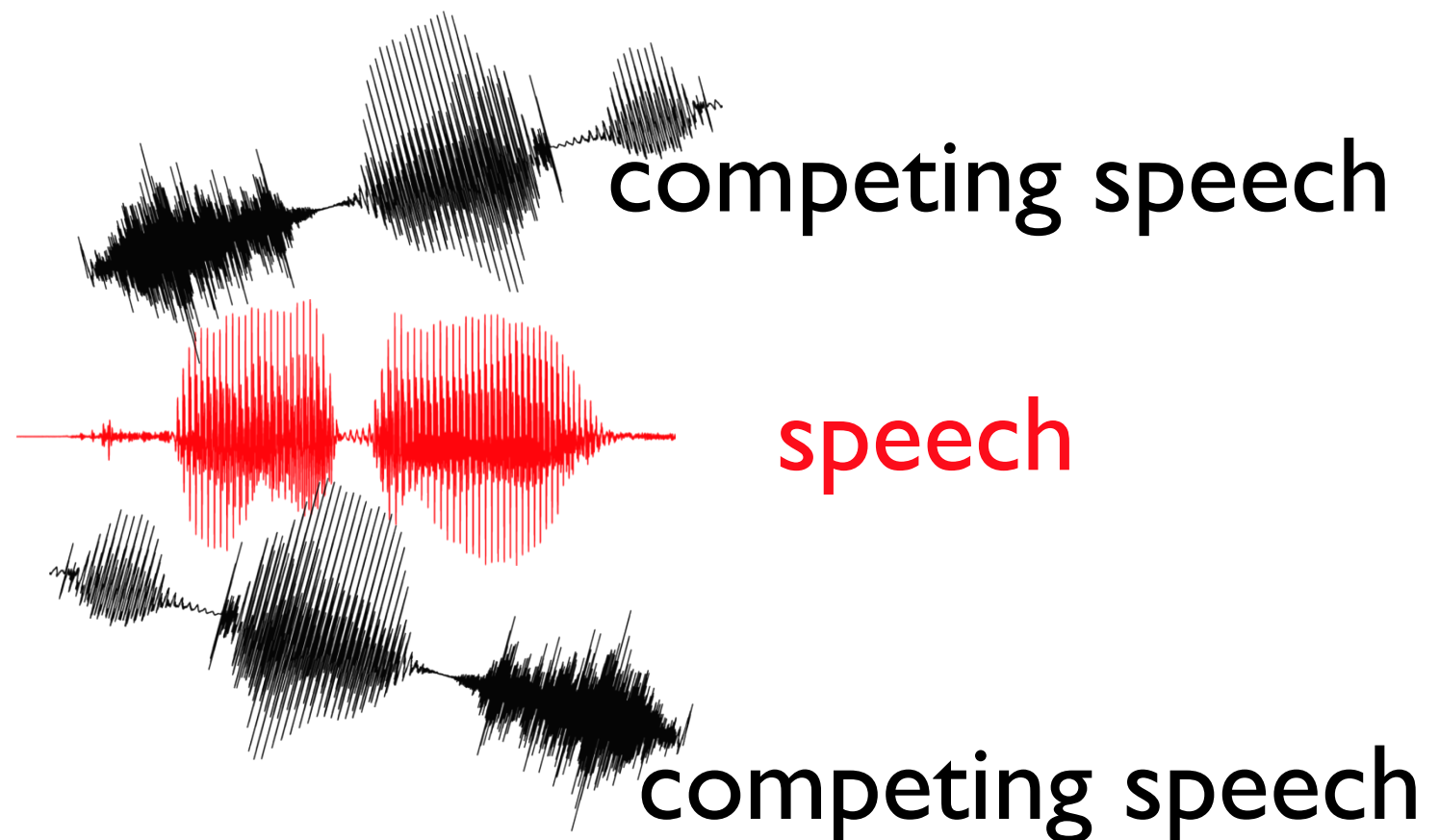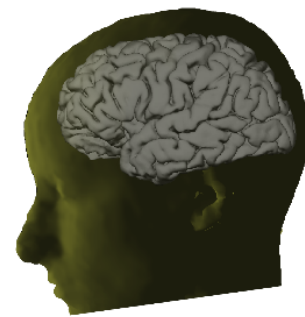
Neural Sources

- M100$_{STRF}$ source near (same as?) M100 source:
  Planum Temporale

- M50$_{STRF}$ source is anterior and medial to M100 (same as M50?):
  Heschl's Gyrus

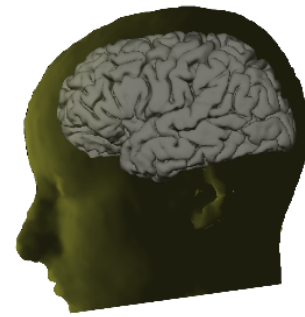- **PT strongly modulated by attention, *but not HG***
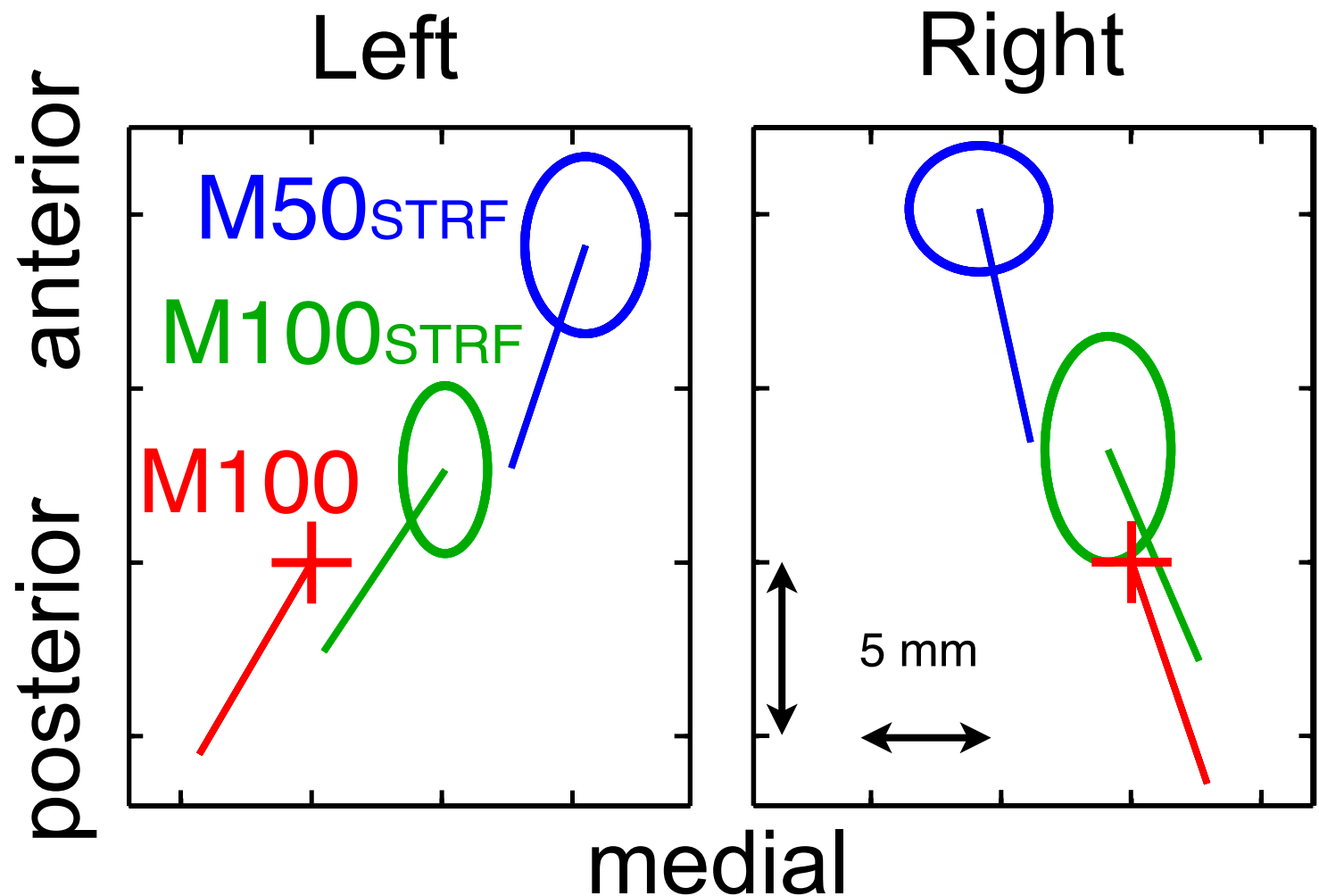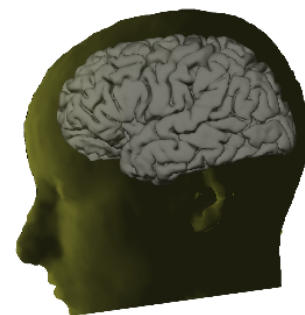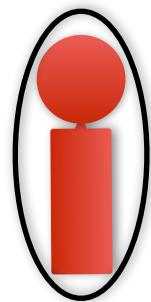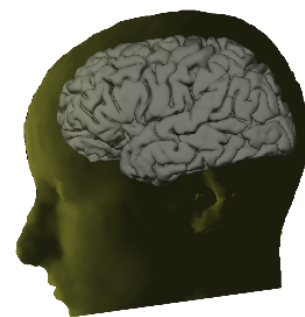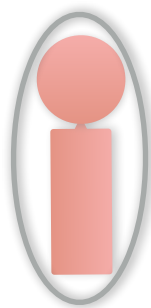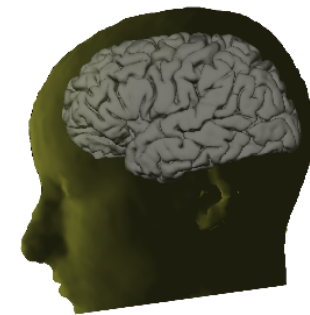
Three Competing Speakers

competing speech

speech

competing speech

Foreground vs. Background
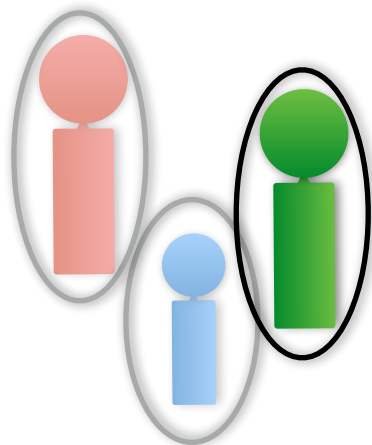
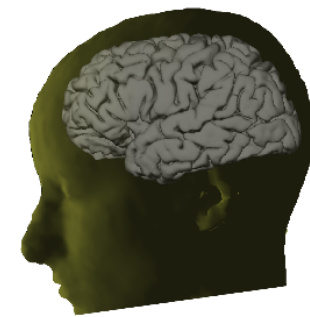# Foreground vs. Background

Foreground vs. Background

Foreground vs. Background

Backgrounds vs. Background

Integration Window over Late Times Only

Backgrounds vs. Background Why not?

Stimulus Background

Speaker 1

Speaker 2

MEG Response

Two Speakers

Backgrounds vs. Background Why not?

Stimulus Background

Speaker 1

Speaker 2

MEG Response

Two Speakers

Backgrounds vs. Background

Individual Speech Streams

reconstruction accuracy

Noise Floor — **Individual Backgrounds** — **Fused Background** — Foreground

Integration Window over Late Times Only

Backgrounds vs. Background
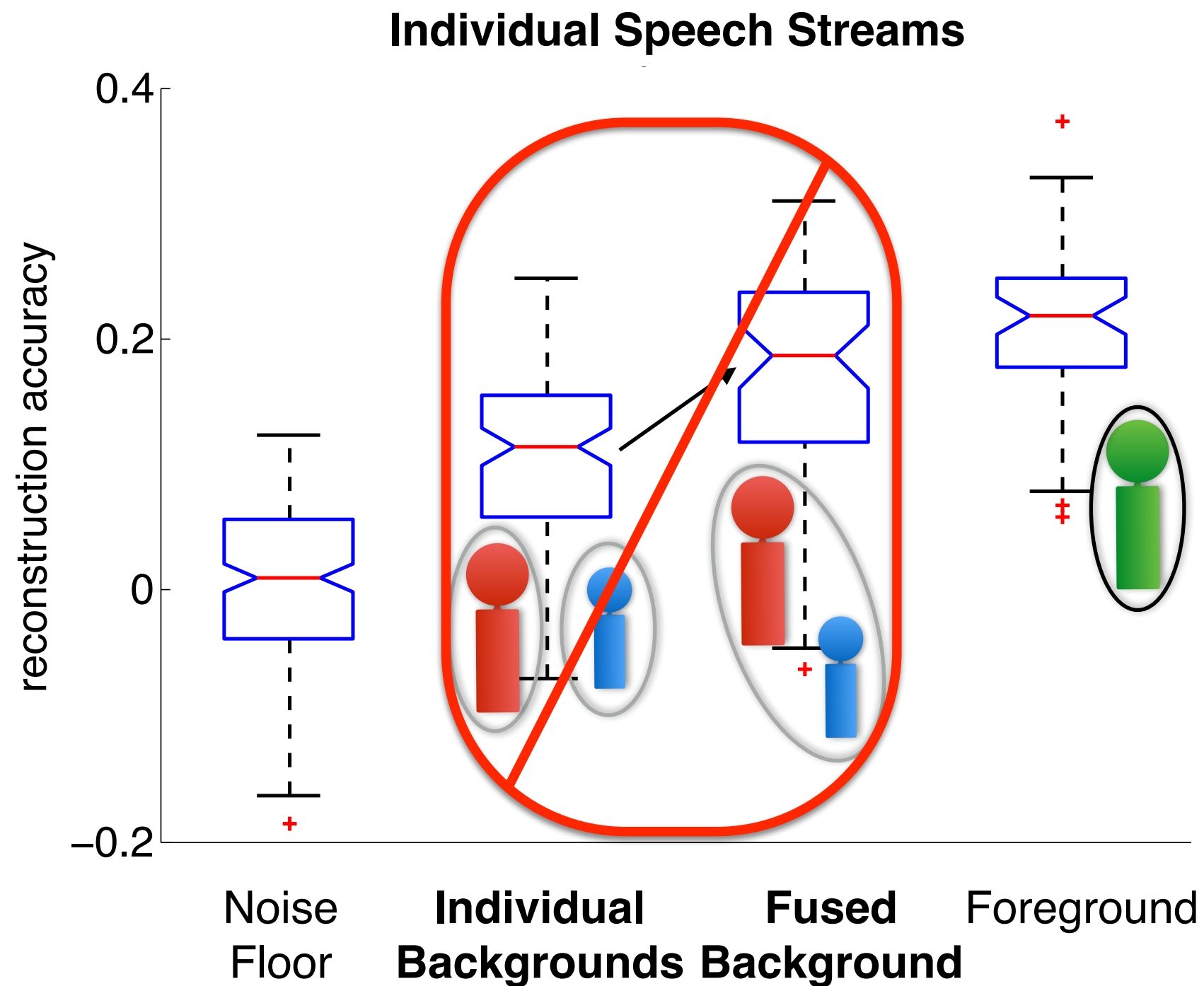
Individual Speech Streams

Integration Window over Late Times Only
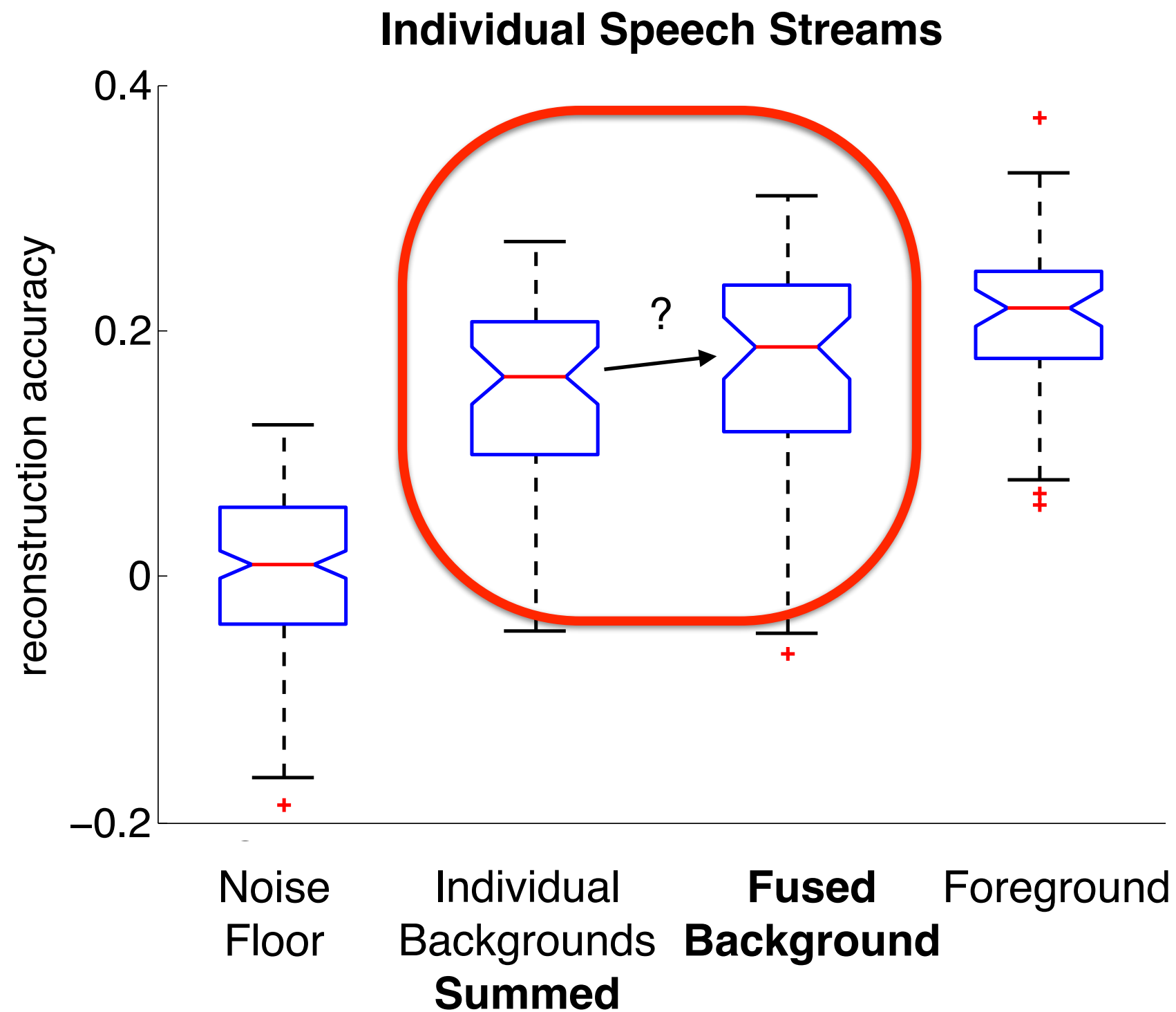
Backgrounds vs. Background

Individual Speech Streams

reconstruction accuracy

0.4

0.2

0

-0.2

?

Noise
Floor

Individual
Backgrounds
Summed
+

Fused
Background

Foreground

Integration Window over Late Times Only

Backgrounds vs. Background

Background Representations

Fused Background

Individual Backgrounds Summed

Integration Window over Late Times Only

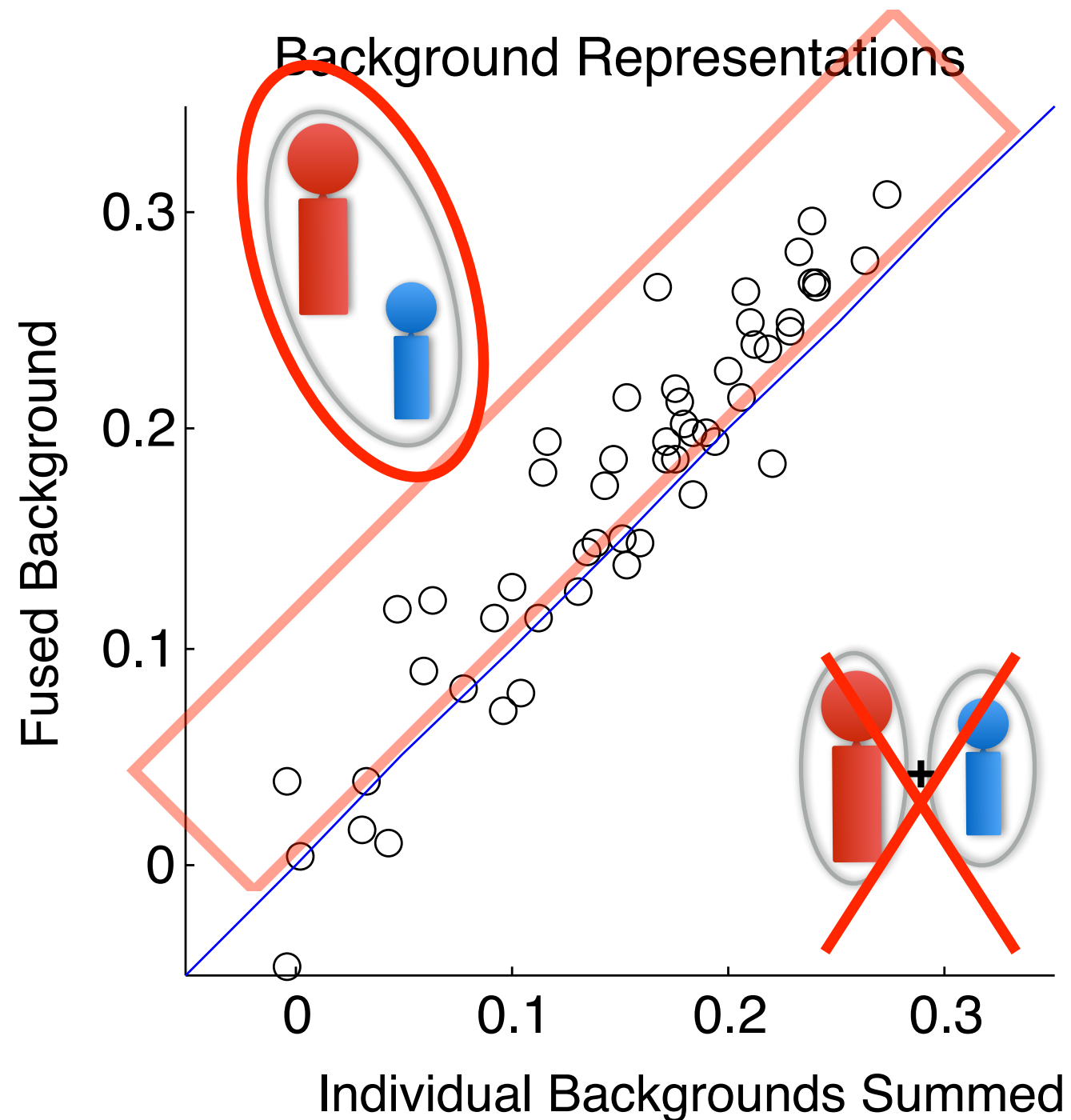# Backgrounds vs. Background
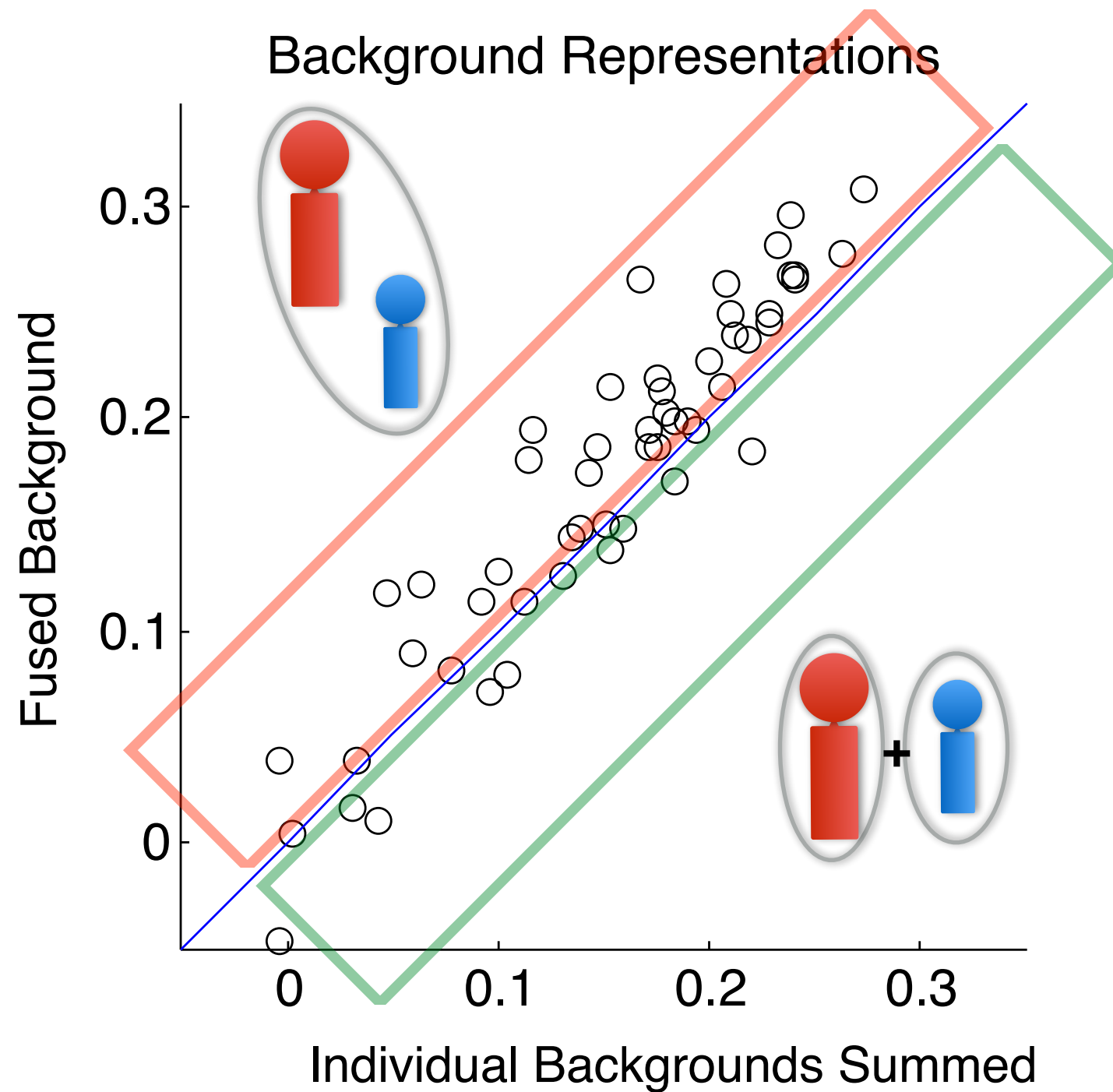
Background Representations
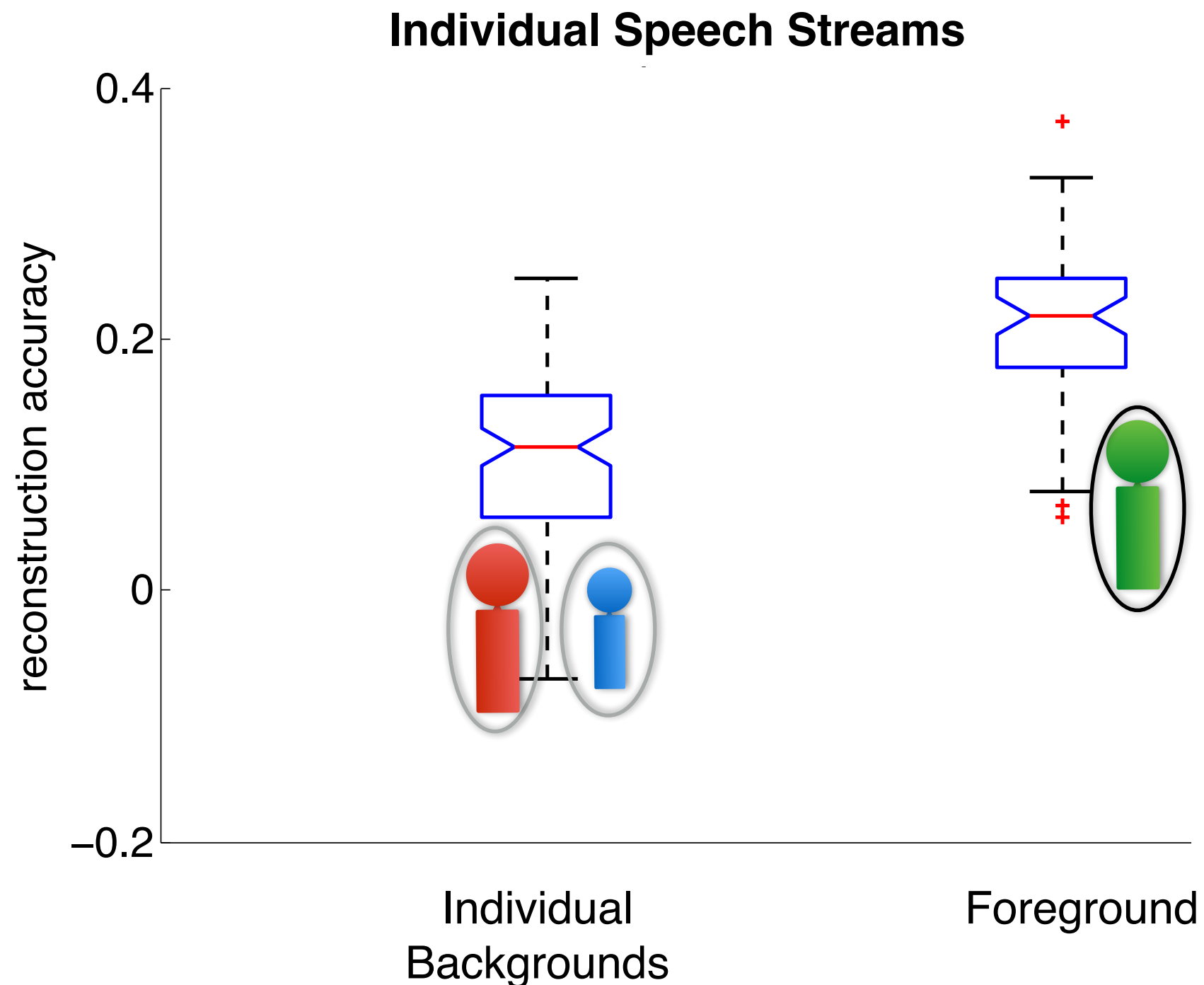
High latency areas (PT) represent *fused* background with better fidelity than *individual* backgrounds (p = 1.3E-05)

Individual Backgrounds Summed

Fused Background

Integration Window over Late Times Only

Foreground vs. Background

Individual Speech Streams

reconstruction accuracy

0.4

0.2

0

−0.2

Individual
Backgrounds

Foreground

Integration Window over Late Times Only

Foreground vs. Background
Early vs. Late

Early (HG)

Late (PT)

$p = 0.03$

$p = 1E-17$
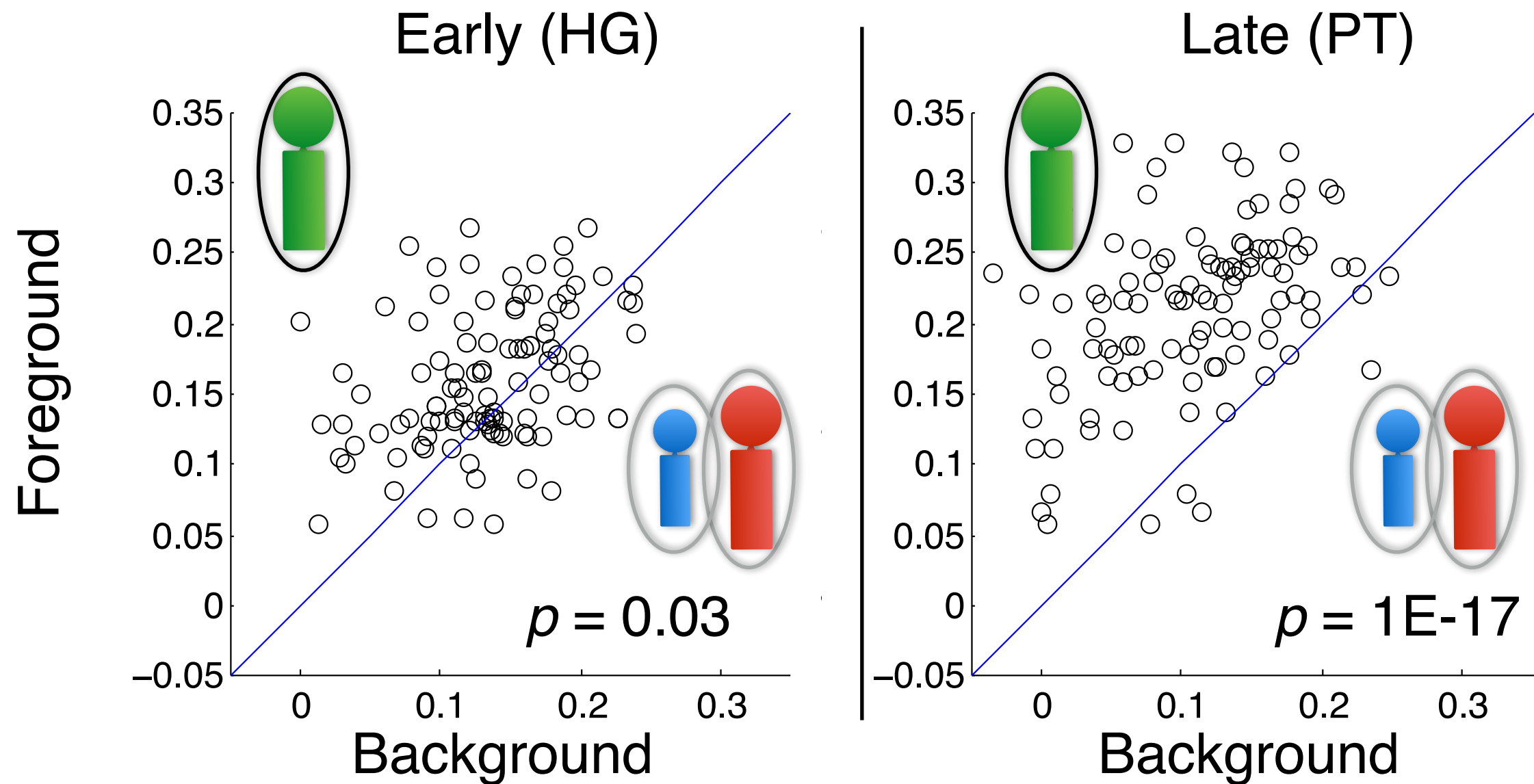
HG represents attended and unattended speech with *almost* equal fidelity

Summary

- Cortical representations of speech
    - ✓ representation of envelope (up to ~10 Hz)
- Object representation at 100 ms latency (PT), but not by 50 ms (HG)
- Consistent with being neural representations of auditory perceptual object
- Preliminary evidence for
    - ✓ PT: additional fused background representation
    - ✓ HG: *almost* equal representations

Thank You