Multisensory Speech Perception: Models and Mechanisms

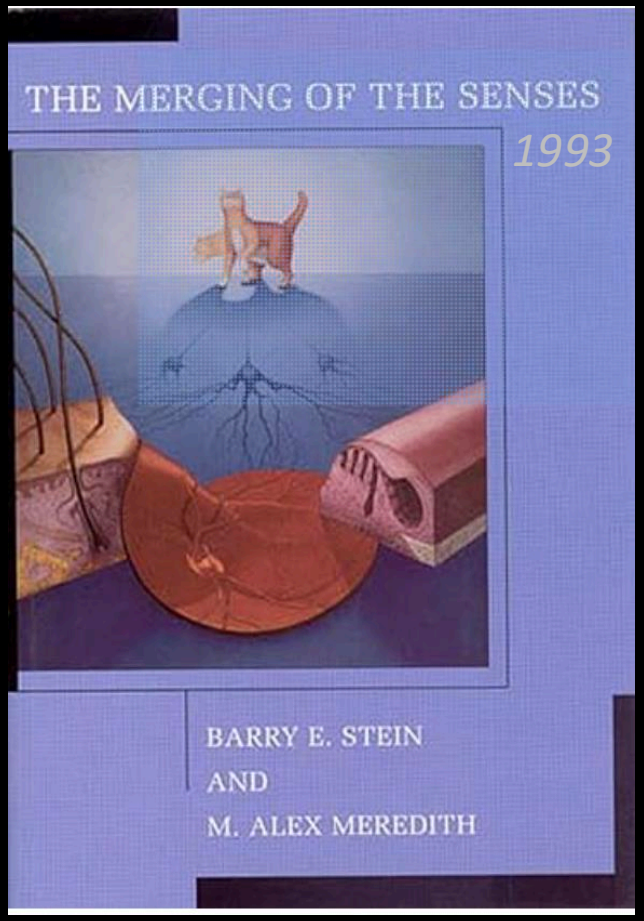
Michael S. Beauchamp, Ph.D.
Professor and Vice Chair for Research
michael.beauchamp@pennmedicine.upenn.edu

data
code
PDFs
beauchamplab.com

Vision Augmented Hearing
Theo Murphy Royal Society Meeting,
Wednesday, March 4, 2026

Multisensory integration

- enhances **speed** and **accuracy** of perception



Human communication is fundamentally multisensory, esp. audiovisual

- enhanced **speed** of processing is critical for the rapid pace of social interactions.
- enhanced **accuracy** is critical since social interactions are high stakes

Beauchamp MS. The Social Mysteries of the Superior Temporal Sulcus. Trends in Cognitive Sciences 2015 Sep;19(9):489-90.

Multisensory speech perception

- **speech content:**
- ***visual*:** viewed face and mouth movements
- ***auditory***: voice

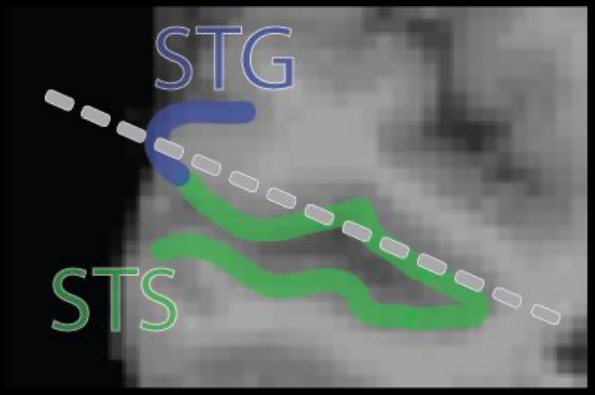
- We learn speech using both the voice and face
- Most studies of language present only one modality (auditory recordings or written text), neglecting visual information from the talker's face

Talk overview

- Behavior
  - *Magnotti et al. (under review)*
- Neural substrates
  - *Zhang et al. (Journal of Neuroscience, 2025)*
- Deep neural network models
  - *Ma et al. (Psychonomic Bulletin and Review, 2026)*

BOLD fMRI

intracranial EEG

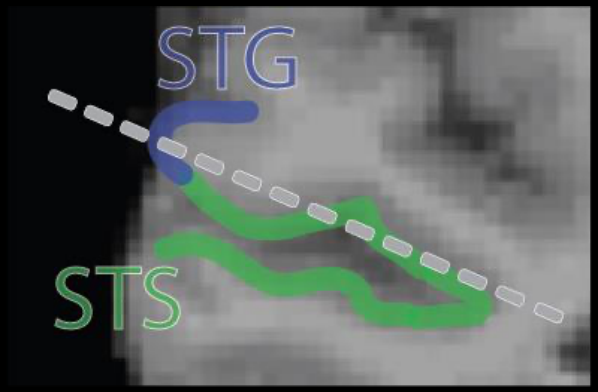


Talk timing

- Behavior
  - 8 min + questions
- Neural substrates
  - 8 min + questions
- Deep neural network models
  - 8 min + questions
- Discussion
  - 15 minutes

BOLD fMRI

intracranial EEG



Noisy speech: the face of the talker serves as a **natural hearing aid**

THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 26, NUMBER 2 MARCH, 1954

Visual Contribution to Speech Intelligibility in Noise*

W. H. Sumby† and Irwin Pollack
Human Factors Operations Research Laboratories, Washington 25, D. C.

(Received November 5, 1953)

John Magnotti

Phoneme overlap analysis of the visual benefit for speech-in-noise perception

John F Magnotti, Yue Zhang, Lin L Zhu, Yingjia Yu, Michael S Beauchamp

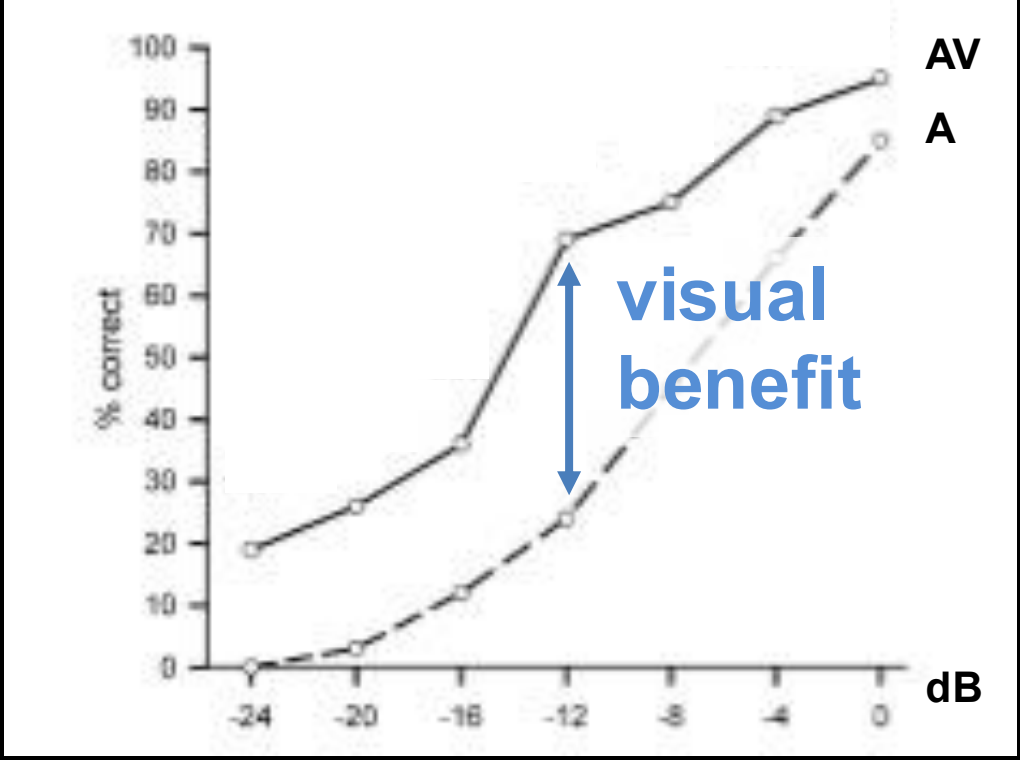
Phil. Trans. R. Soc. B

(under review)

Noisy auditory word

Noisy audiovisual word

The visual benefit for speech-in-noise perception



Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments

Lars A. Ross[1,2], Dave Saint-Amour[2], Victoria M. Leavitt[2,3], Daniel C. Javitt[1,2] and John J. Foxe[1,2,3]

Cerebral Cortex May 2007;17:1147-1153
doi:10.1093/cercor/bhl024
Advance Access publication June 19, 2006

Conventional measure

binary, whole-word

correct (100%)*: response matches stimulus word*

incorrect (0%)*: response does not match stimulus*

Sumby and Pollack (1954)
Ross, Saint-Amout, Leavitt, Javitt, Foxe (2007)
Fiscella, Cappelloni, Maddox (2022)
Alampounti, Cooper, Rosen, Bizley (2026)
+ *many more*

Conventional measure

correct (100%)*: response matches stimulus word*

incorrect (0%)*: response does not match stimulus*

- Stimulus word: "theme"

Conventional measure

correct (100%)*: response matches stimulus word*

incorrect (0%)*: response does not match stimulus*

- Stimulus word: "theme"
- Present in auditory (A) format
  - response: "blue"

Conventional measure

correct (100%)*: response matches stimulus word*

incorrect (0%)*: response does not match stimulus*

- Stimulus word: "theme"
- Present in auditory (A) format
  - response: "blue" X 0%

Conventional measure

correct (100%)*: response matches stimulus word*

incorrect (0%)*: response does not match stimulus*

- Stimulus word: "theme"
- Present in auditory (A) format
  - response: "blue" X 0%
- Present in audiovisual (AV) format
  - response: "thief" X 0%

Conventional measure

correct (100%)*: response matches stimulus word*

incorrect (0%)*: response does not match stimulus*

- Stimulus word: "theme"
- Present in auditory (A) format
  - response: "blue" X 0%
- Present in audiovisual (AV) format
  - response: "thief" X 0%
- Visual benefit: AV 0% - A 0% = 0%

Problem with conventional measure

Stimulus word: "**the**me"

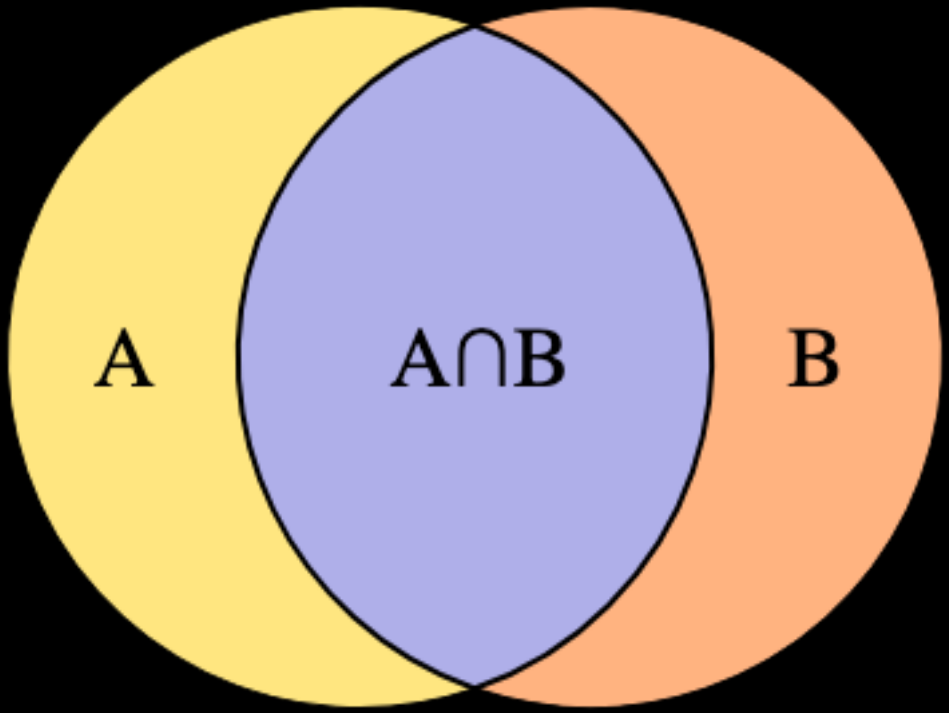
auditory format: "blue"

audiovisual format: "**thie**f"

No visual benefit ?

Alternative: phoneme-overlap

Jaccard index: intersection / union



phoneme-overlap

stimulus: Ph1 Ph2 Ph3

response: Ph1 Ph4 Ph5

phoneme-overlap

stimulus: Ph1 Ph2 Ph3

response: Ph1 Ph4 Ph5

union = 5 (Ph1 Ph2 Ph3 Ph4 Ph5)

intersection = 1 (Ph 1)

phoneme-overlap

stimulus: Ph1 Ph2 Ph3

response: Ph1 Ph4 Ph5

union = 5 (Ph1 Ph2 Ph3 Ph4 Ph5)

intersection = 1 (Ph 1)

Jaccard index = intersection / union = 1 / 5

Phoneme-overlap = 20%

phoneme-overlap

Stimulus word: "theme" ***TH IY M***

A format response: "blue" ***BL U W***

A union = 6, intersection = 0, accuracy = 0%

phoneme-overlap

Stimulus word: "theme" ***TH IY M***

AV format response: "thief" ***TH IY F***

AV union = 4, intersection = 2, accuracy = 50%

phoneme-overlap

Stimulus word: "theme" ***TH IY M***

A union = 6, intersection = 0, accuracy = 0%

AV union = 4, intersection = 2, accuracy = 50%

visual benefit: ***AV - A*** = 50% - 0% = **50%**

phoneme-overlap

Stimulus word: "theme" ***TH IY M***

A format response: "blue" ***BL U W***

AV format response: "thief" ***TH IY F***

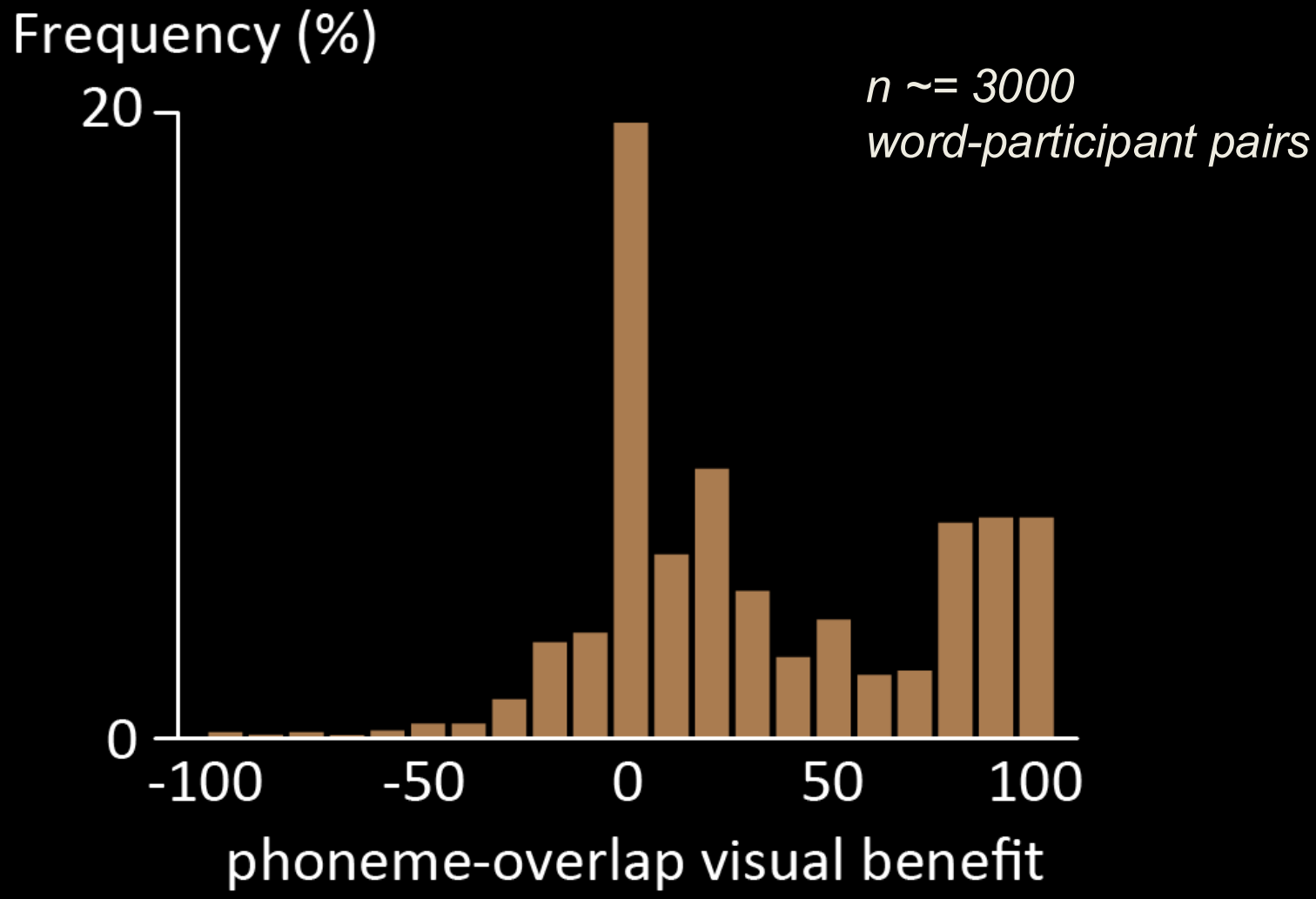
phoneme-overlap visual benefit: 50% - 0% = **50%**

vs. whole-word visual benefit: **0%**

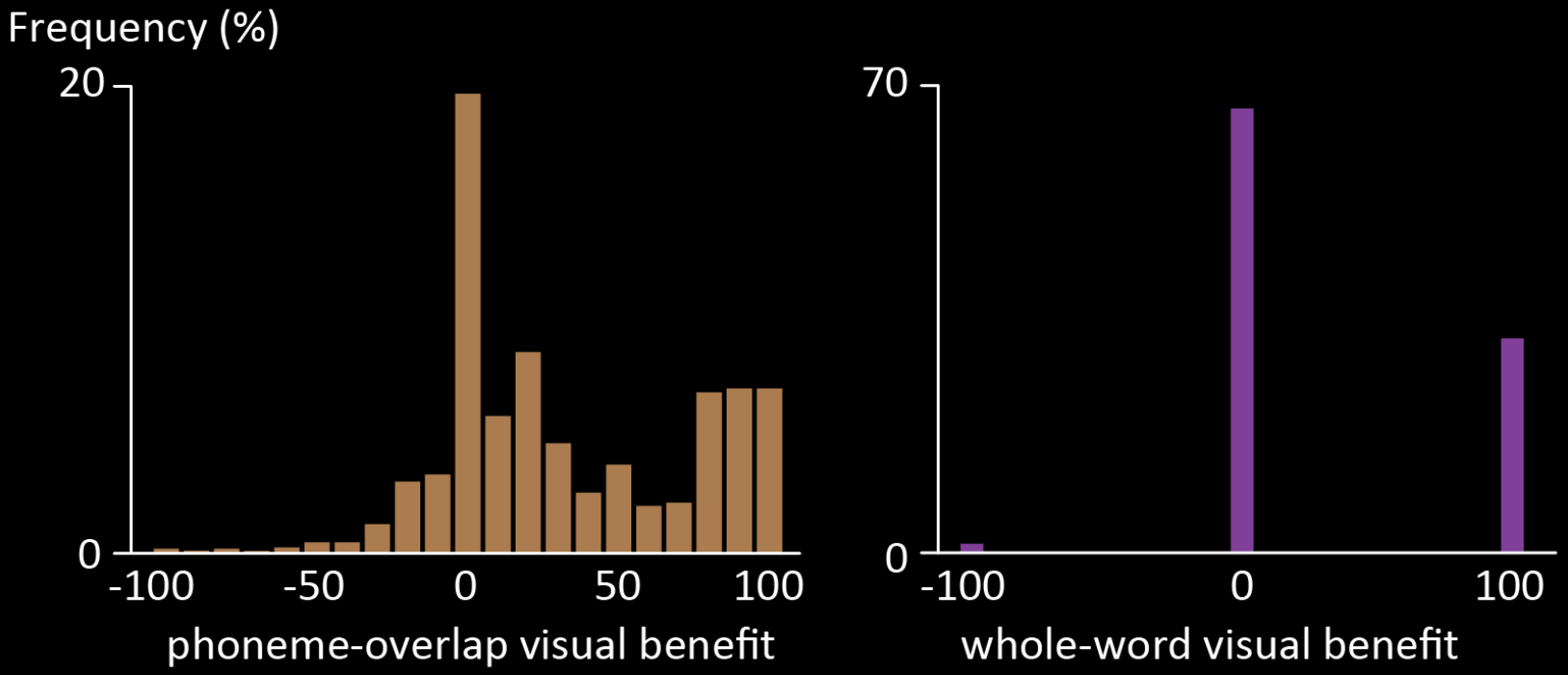
Methods

- online testing: 55 participants, up to 110 words presented to each participant
- -12 dB auditory (speech-shaped) noise, intermediate noise level for maximum visual benefit
- auditory format of a word always presented before audiovisual format (at least 3 intervening trials)

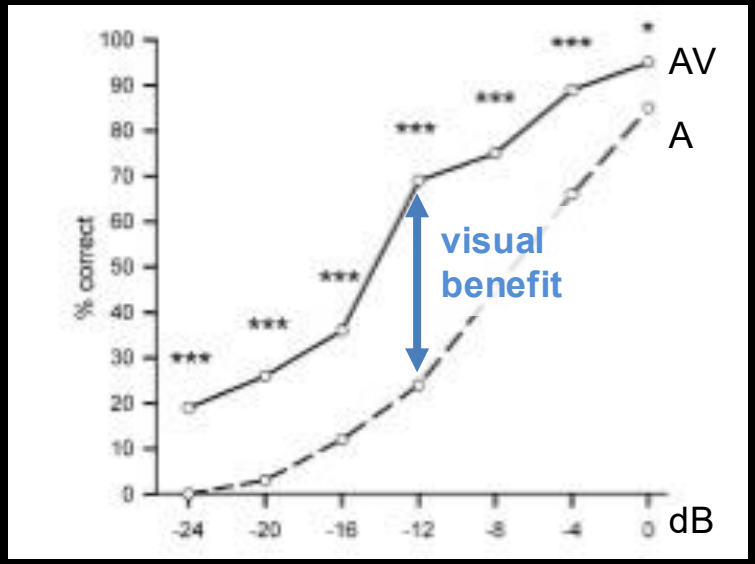
Phoneme-overlap results



Phoneme-overlap *vs.* whole-word

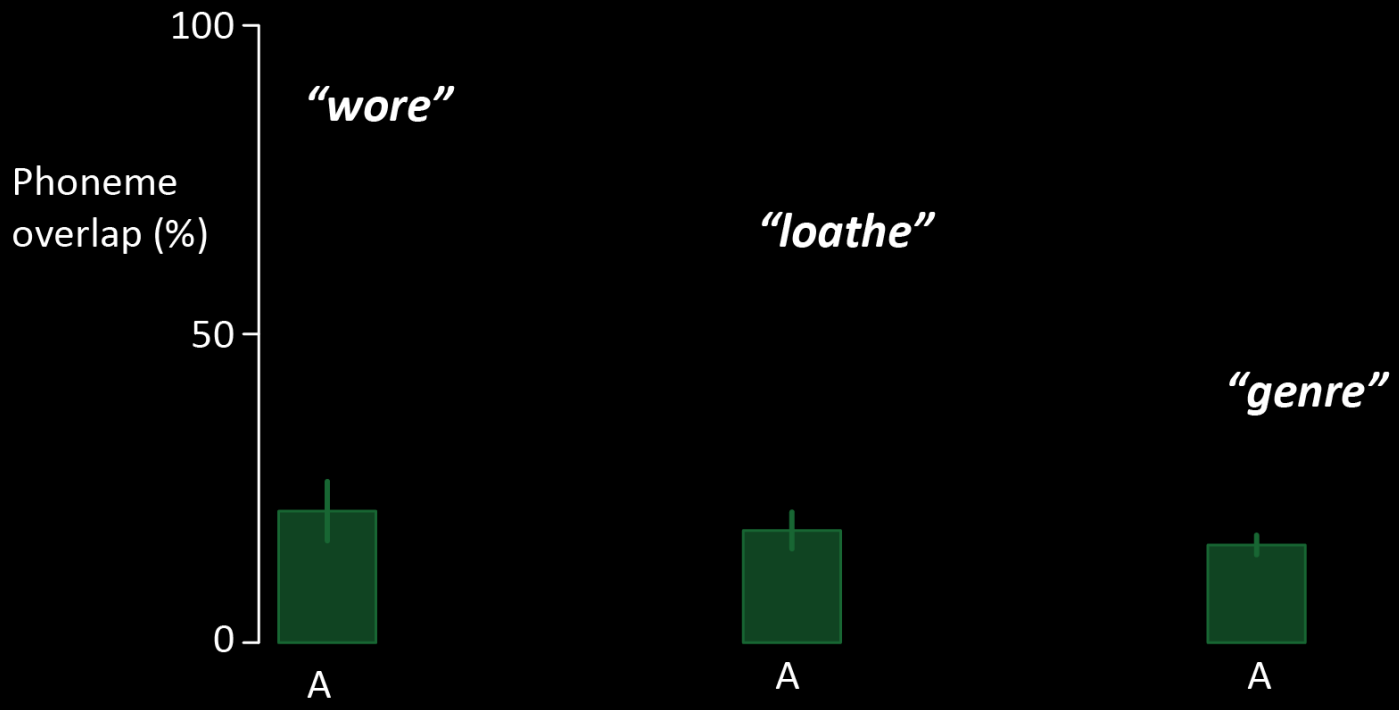


Whole-word: words at a given noise level assumed to have same visual benefit

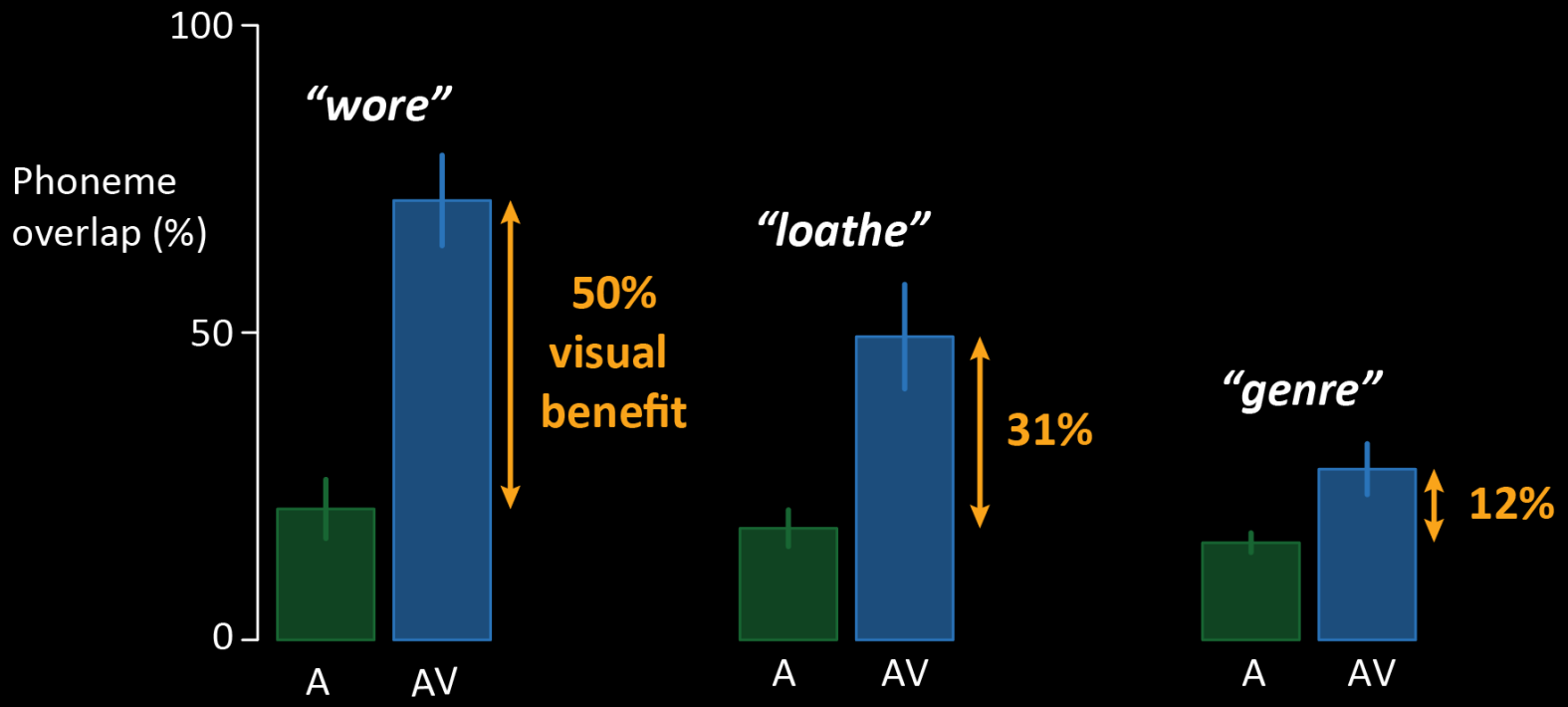


Is this assumption warranted?

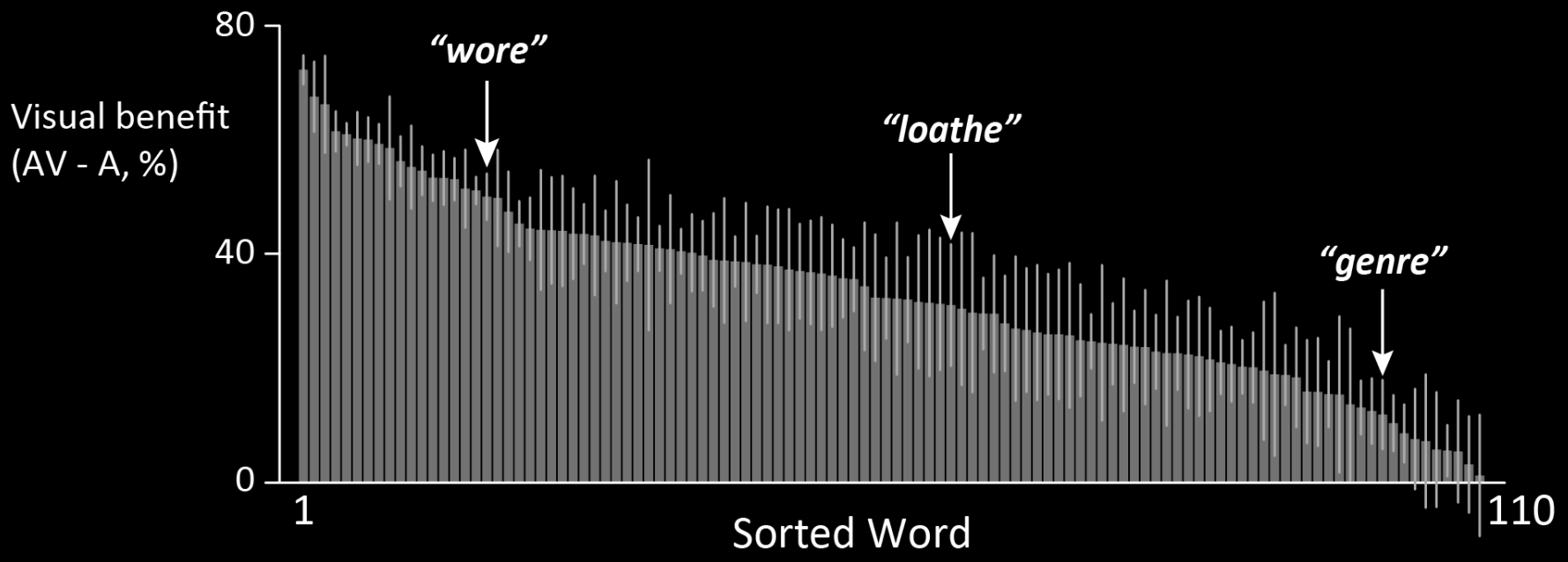
Phoneme-level measurement



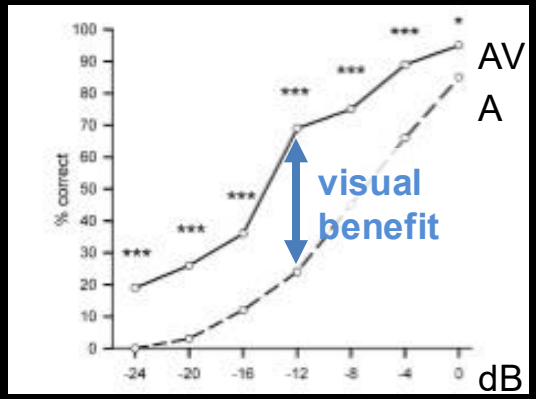
Phoneme-level measurement

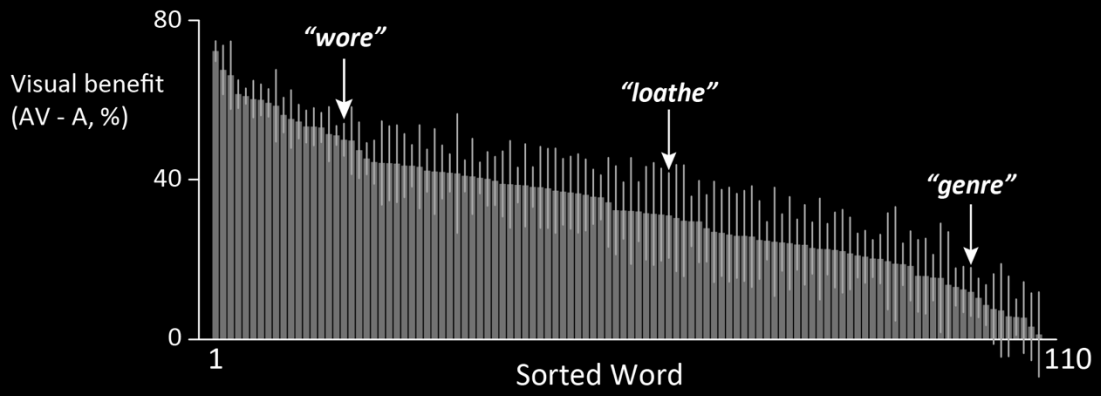


Phoneme-level measurement



Whole-word: all words at given noise level assumed to have same visual benefit

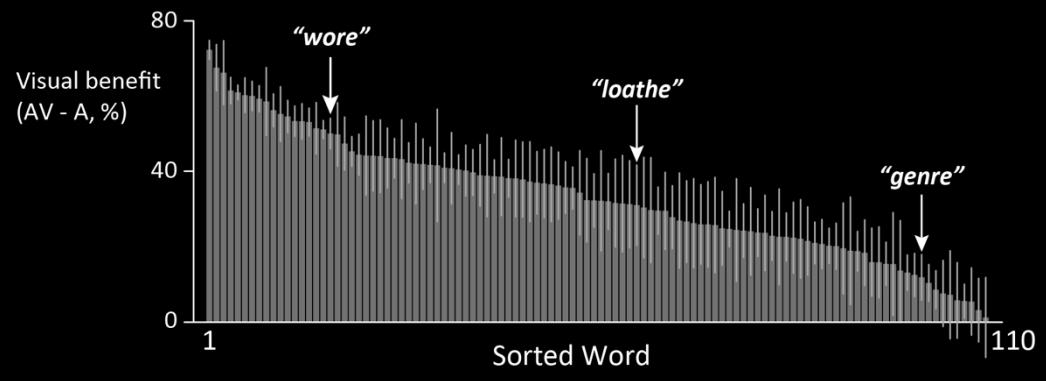




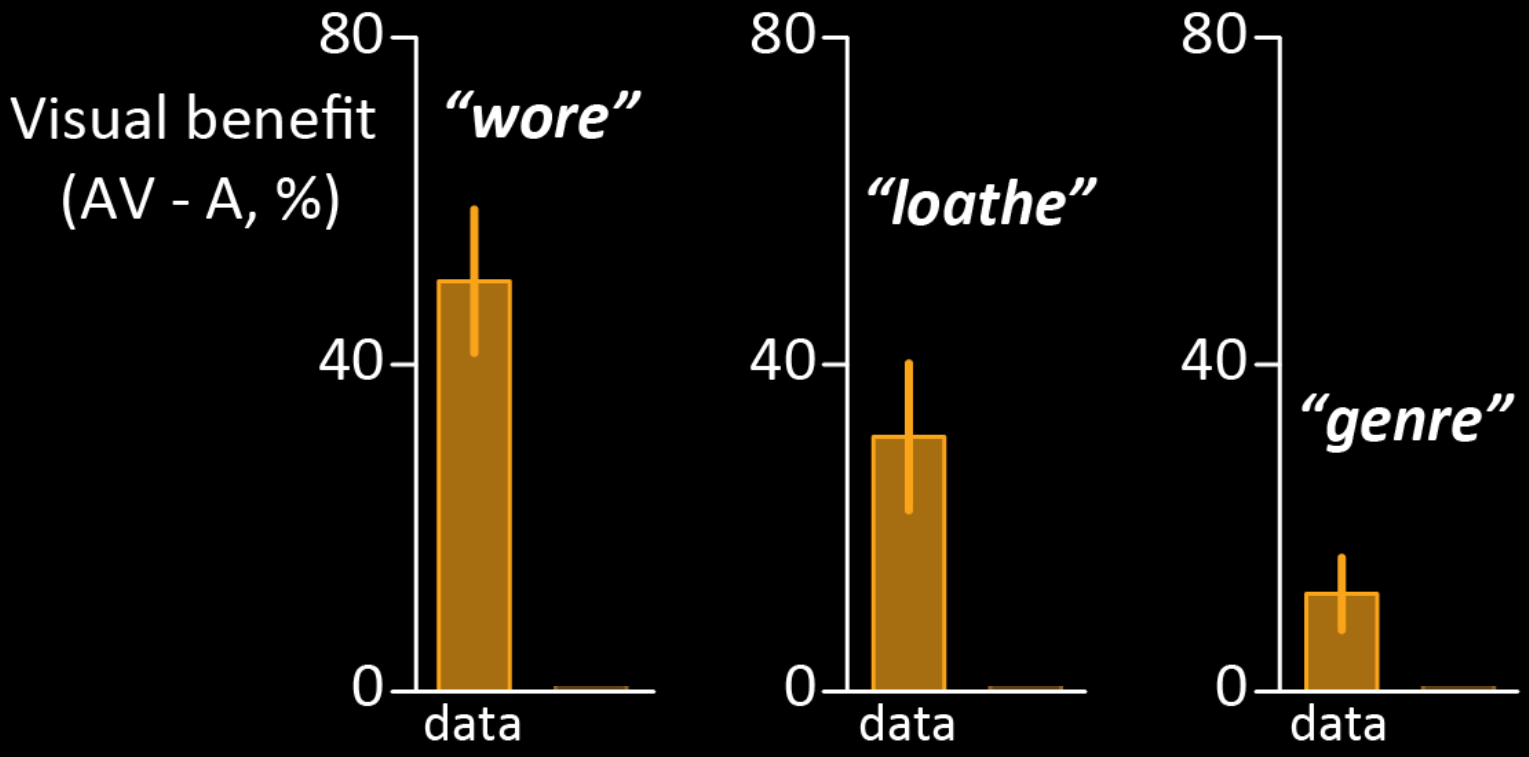
Is this assumption warranted? No!

Phoneme-overlap analysis

- whole-word analysis: all words at given noise level assumed to have same visual benefit
- not so: dramatic variation in visual benefit
- is this just random variation? Or can it be predicted based on phoneme composition?

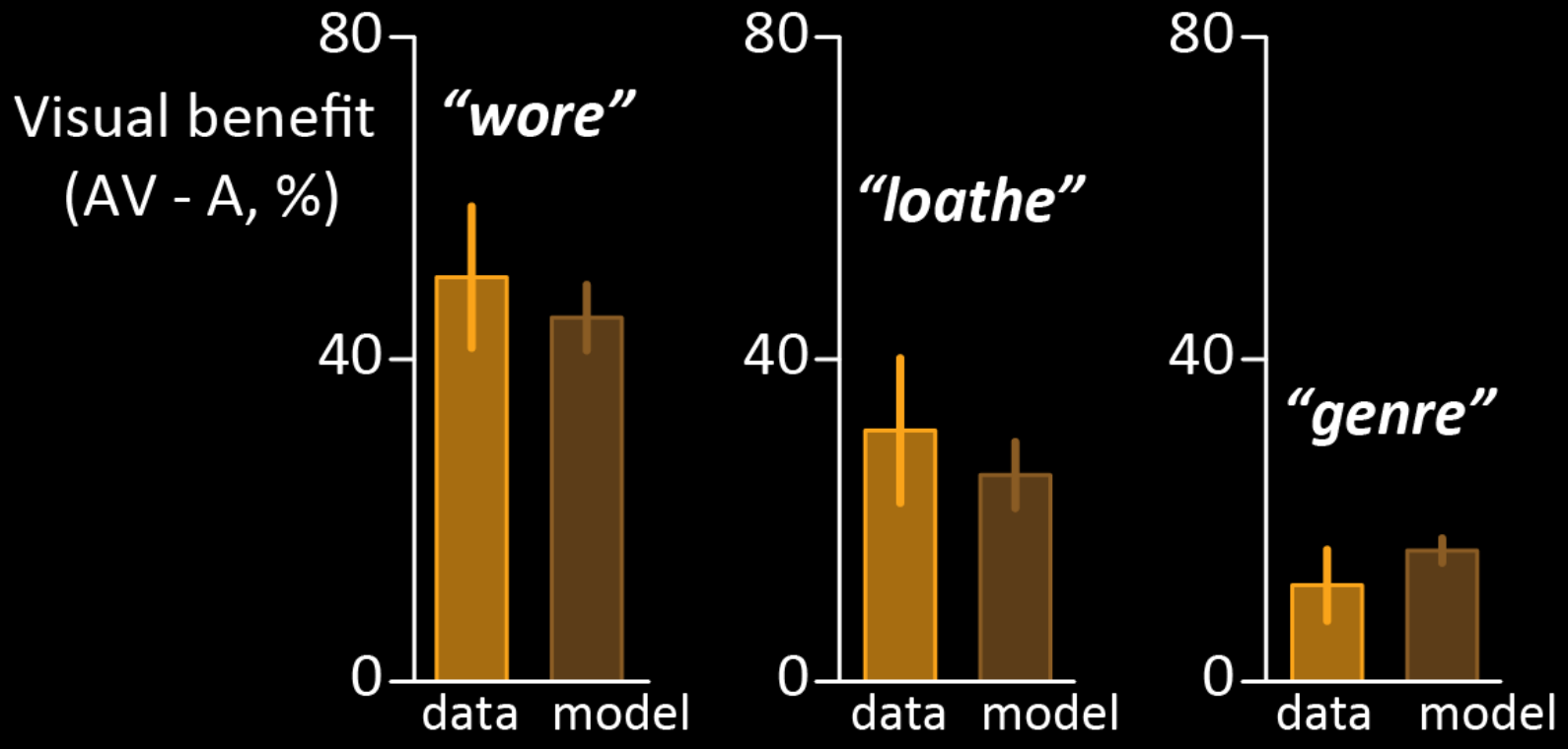


Actual



- Create models with individual words held-out

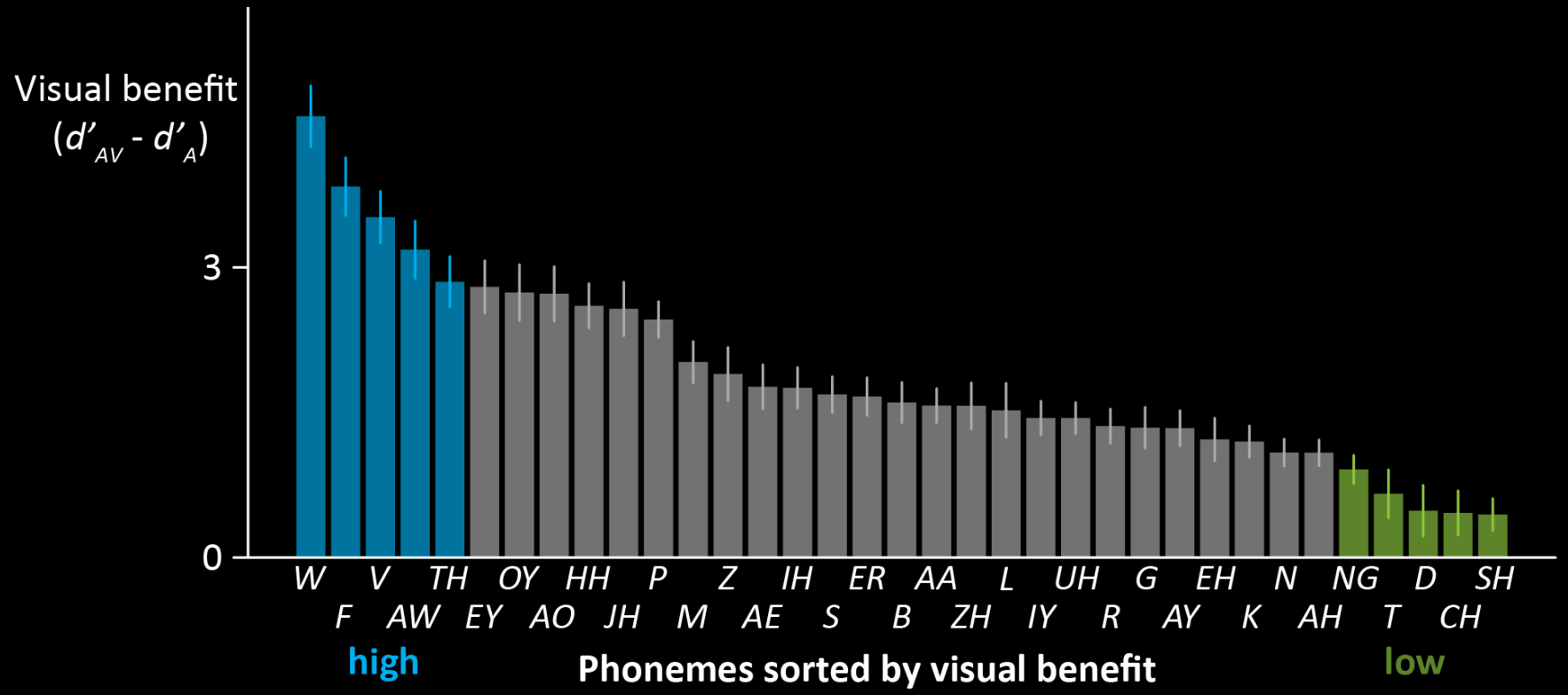
Actual *vs.* predicted



What about individual phonemes?

- inverse analysis: instead of analyzing at word level, analyze individual phonemes
- many words, but few phonemes, so use signal detection theory to calculate sensitivity (*d'*), accounting for false alarms

Phoneme-level sensitivity

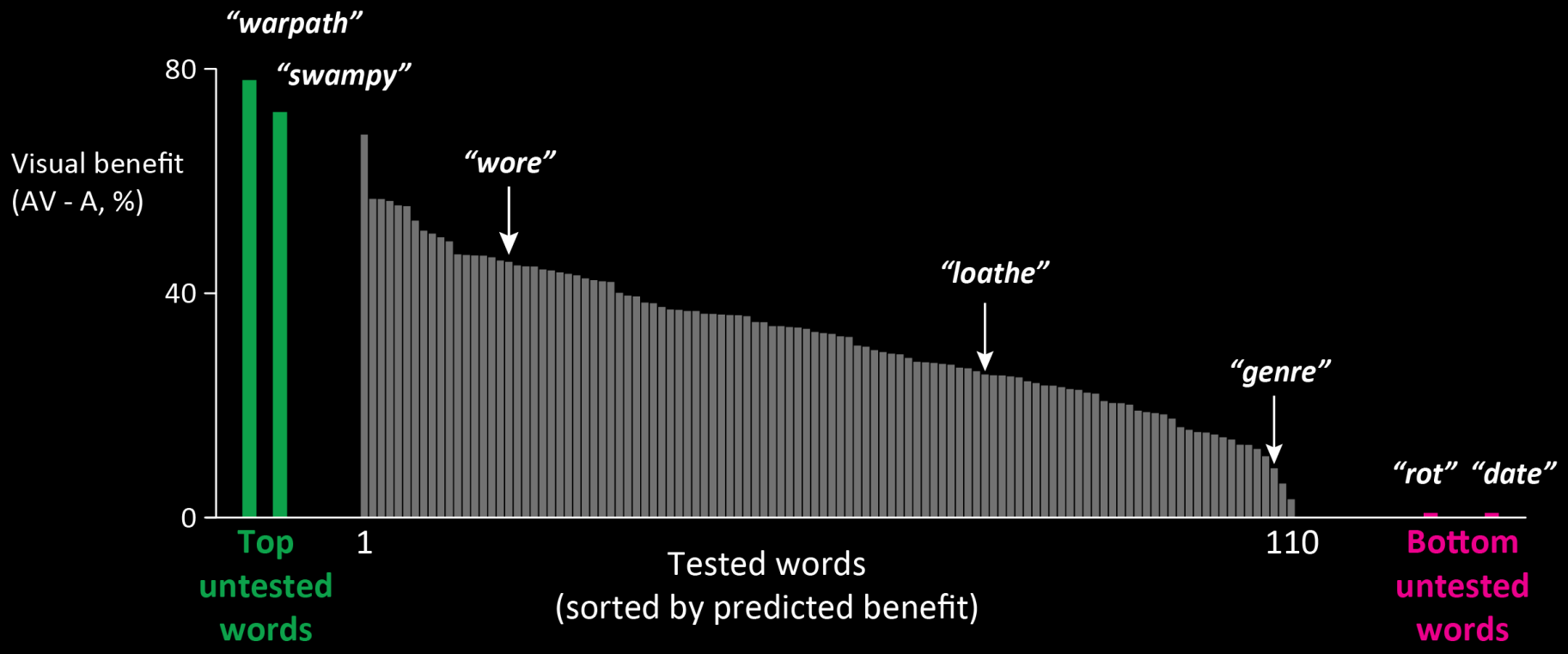


Consistent with viseme classification systems

Jeffers and Barley (1971); Cappelletta & Harte (2011); Bear and Harvey (2017)

Nidiffer *et al.* (2023)

Key advantage: prediction



Section summary

- conventional binary, whole-word analysis is a relatively insensitive measure of multisensory integration; no predictive abilities
- phoneme-level analysis is more sensitive and allows for prediction of untested words
- could be useful for screening or clinical testing
  - choose words with high visual benefit
  - use words from any dictionary (*e.g.* age-normed)

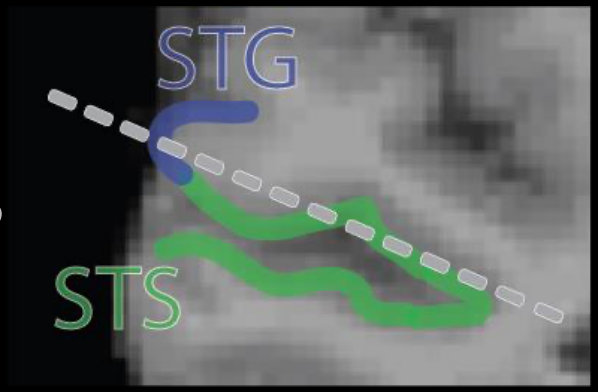
Questions?

Talk Overview

- Introduction to audiovisual speech perception
- Behavior
  - *Magnotti et al. (under review)*
- Neural substrates
  - *Zhang et al. (Journal of Neuroscience, 2025)*
- Deep neural network models
  - *Ma et al. (Psychonomic Bulletin and Review, 2026)*

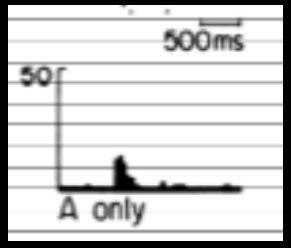
BOLD fMRI

intracranial EEG

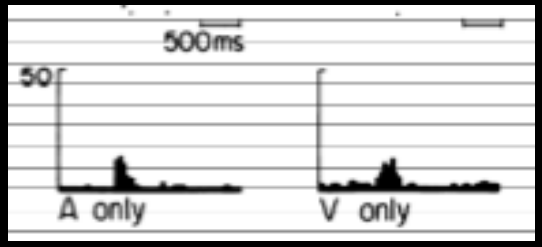


Neural signatures of audiovisual integration: #1: ***Larger*** audiovisual response

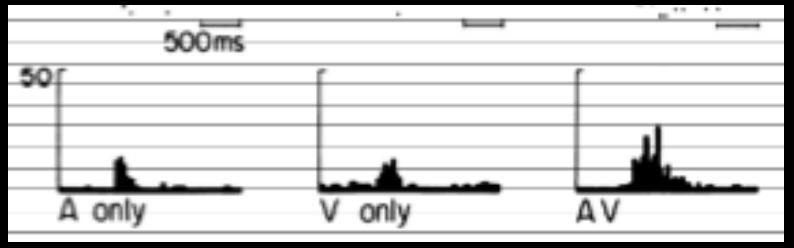
- unisensory response to auditory (A)



- unisensory response to visual (V)

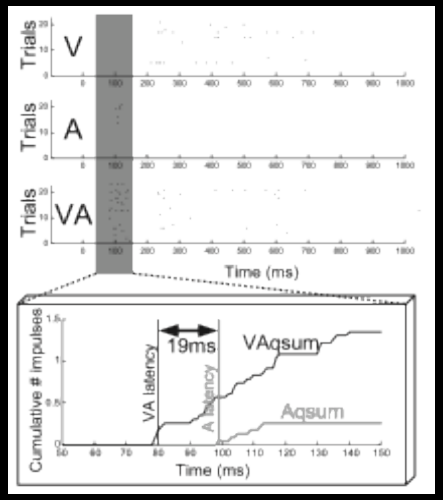


- larger multisensory response (AV)



Meredith, A.M., and Stein, B.E. Journal of Neurophysiology (1986).

Neural signatures of audiovisual integration: #2: ***Faster*** audiovisual response



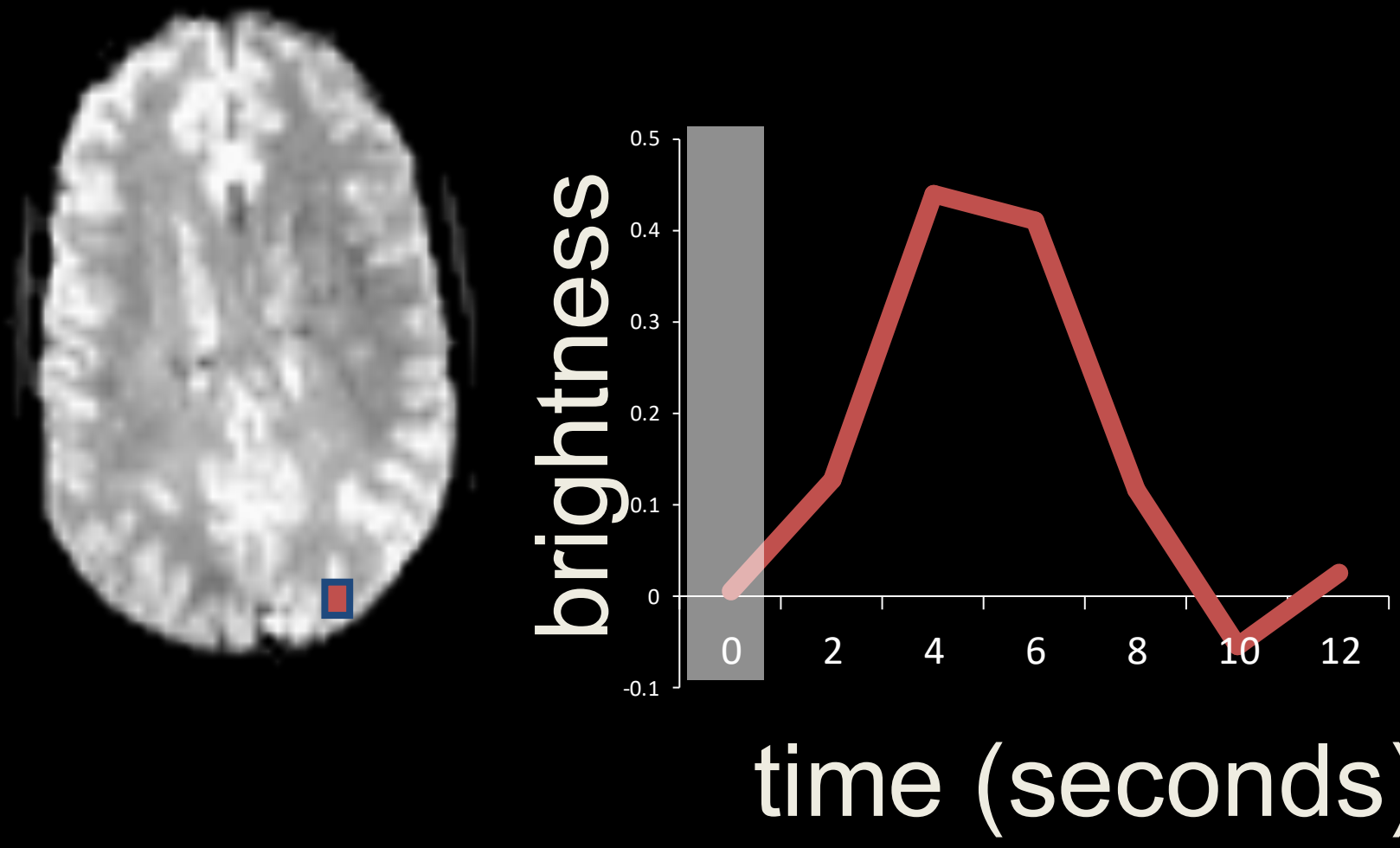
Rowland, B.A., Quessy, S., Stanford, T.R., and Stein, B.E. Journal of Neuroscience (2007).

shorter latency

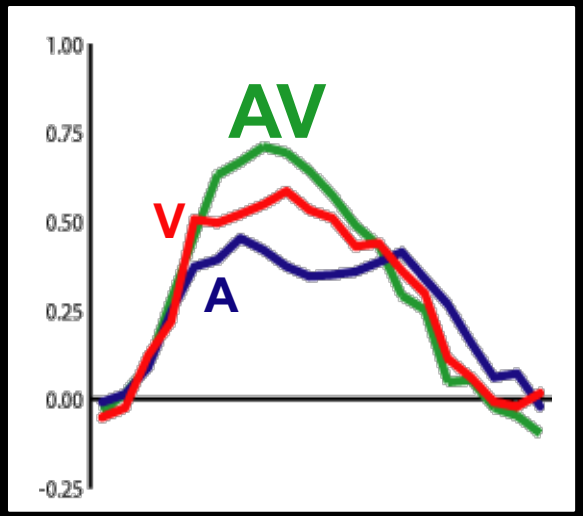
(faster) audiovisual response

3 tesla MR scanner

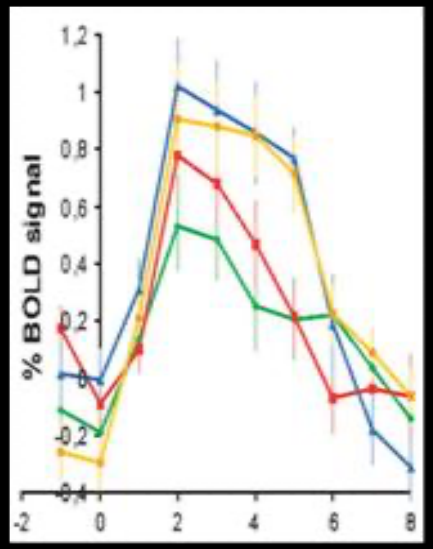
BOLD MR signal



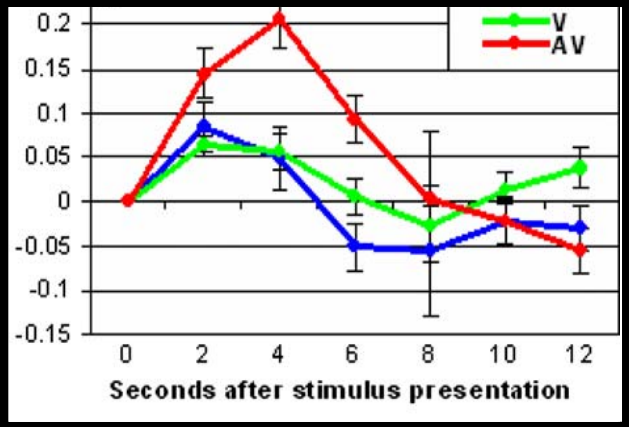
fMRI studies of audiovisual integration: Superior Temporal Sulcus (STS)



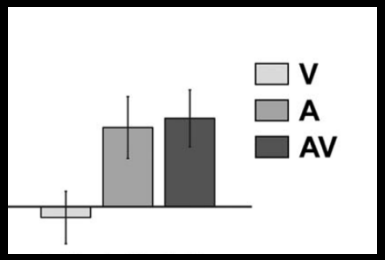
Beauchamp *et al*., 2004



van Atteveldt *et al*., 2007

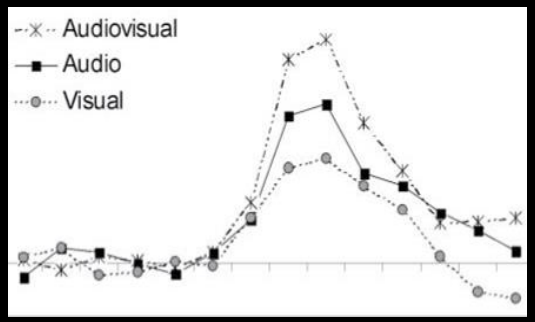


Stevenson *et al*., 2007



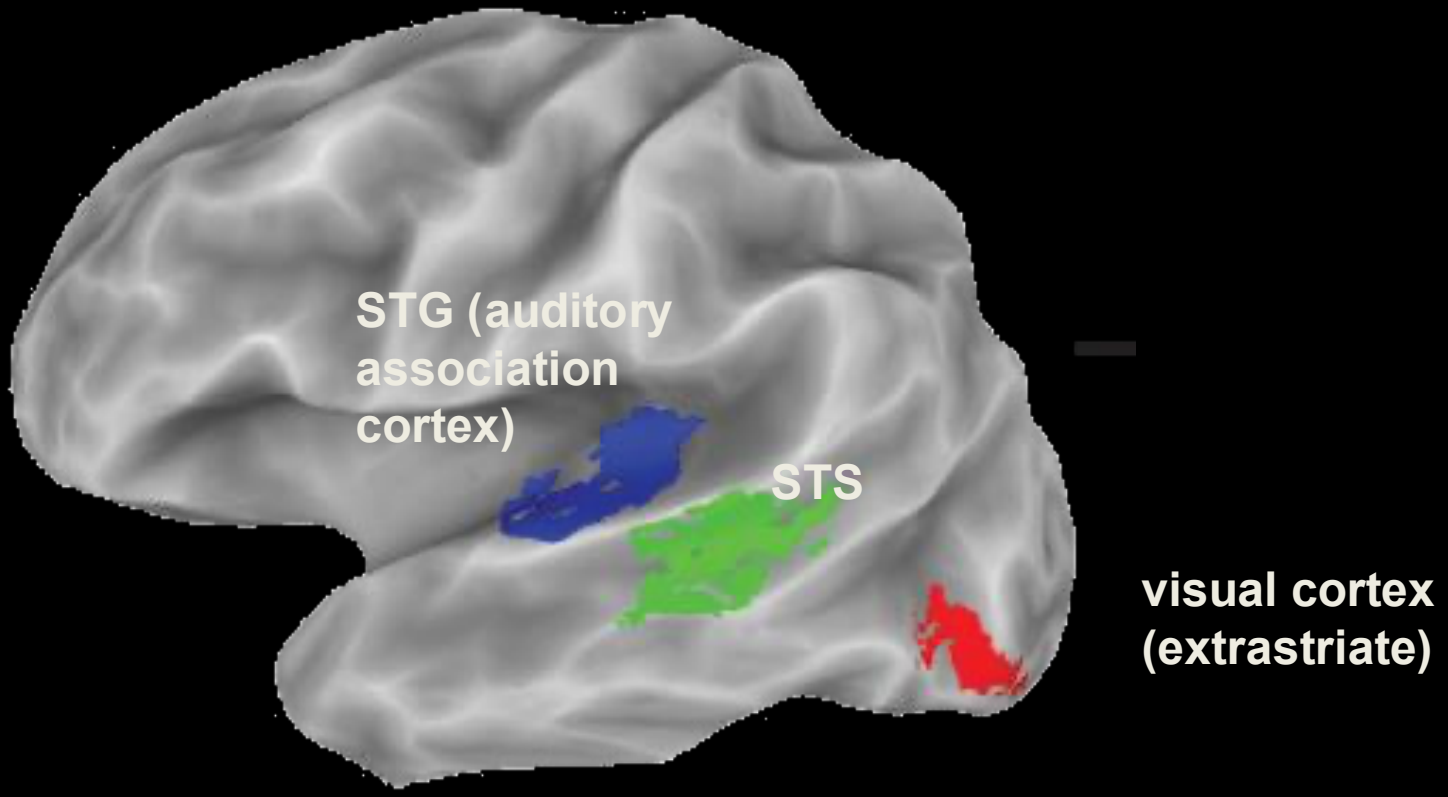
Werner & Noppeney, 2010

+ NHPs



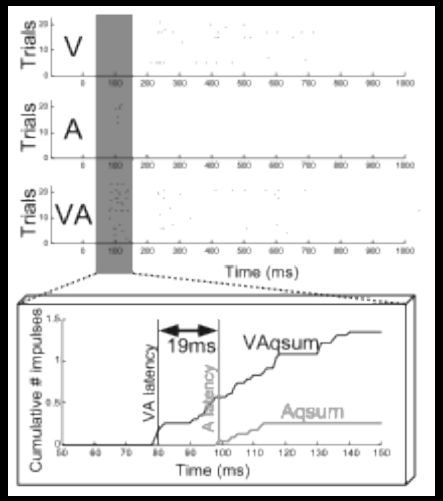
Wright *et al*., 2003

Key areas for audiovisual speech perception



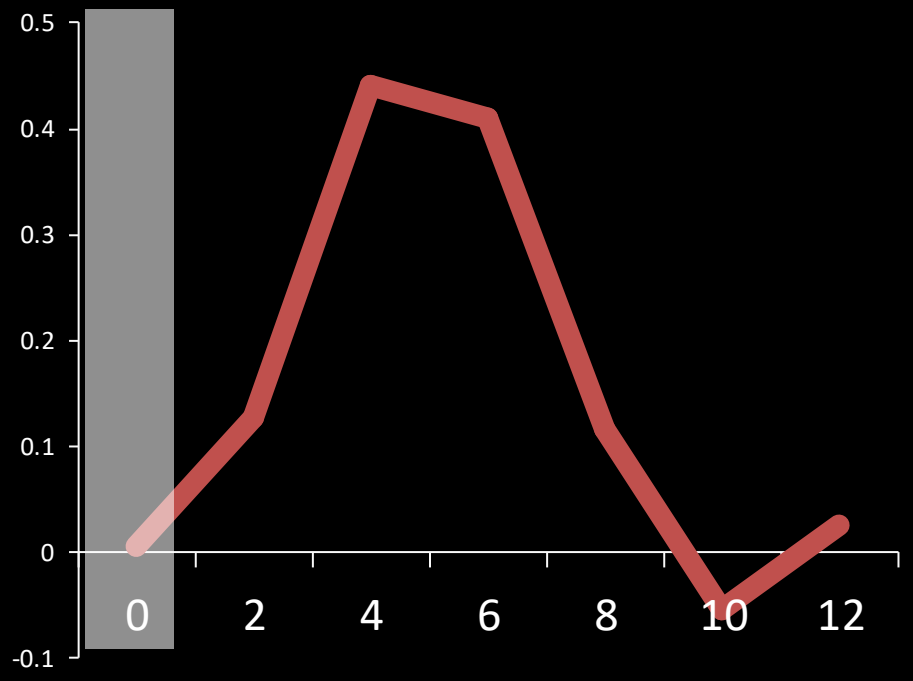
Nath, A. R. & Beauchamp, M.S.. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, 2011

Problem: fMRI is slow and indirect



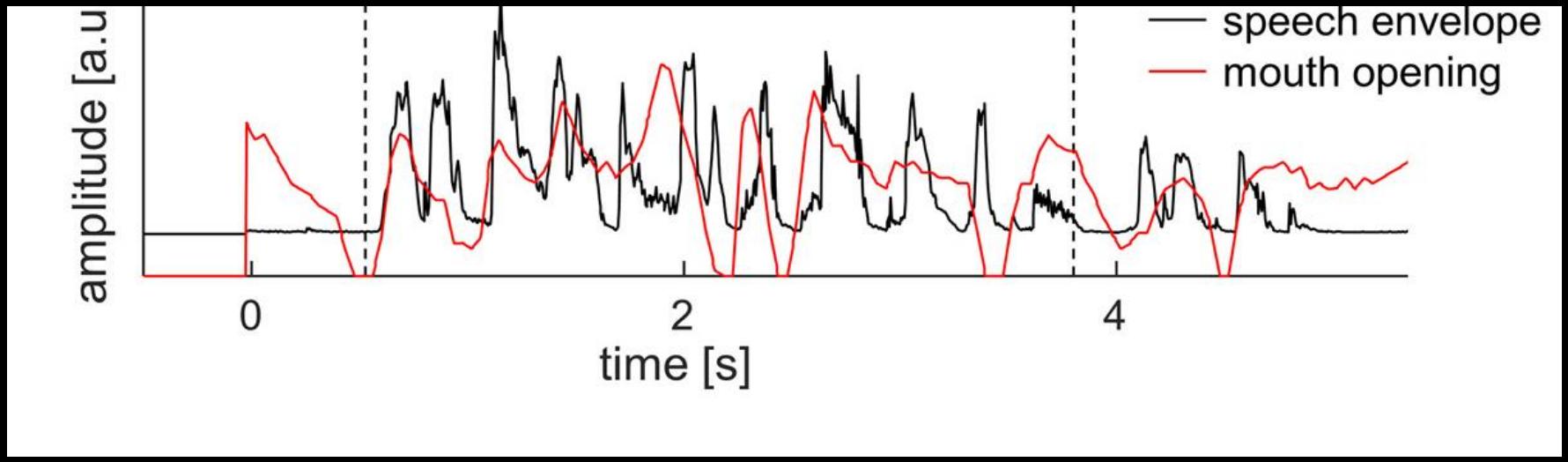
Rowland, B.A., Quessy, S., Stanford, T.R., and Stein, B.E. Journal of Neuroscience (2007).

19 ms shorter latency



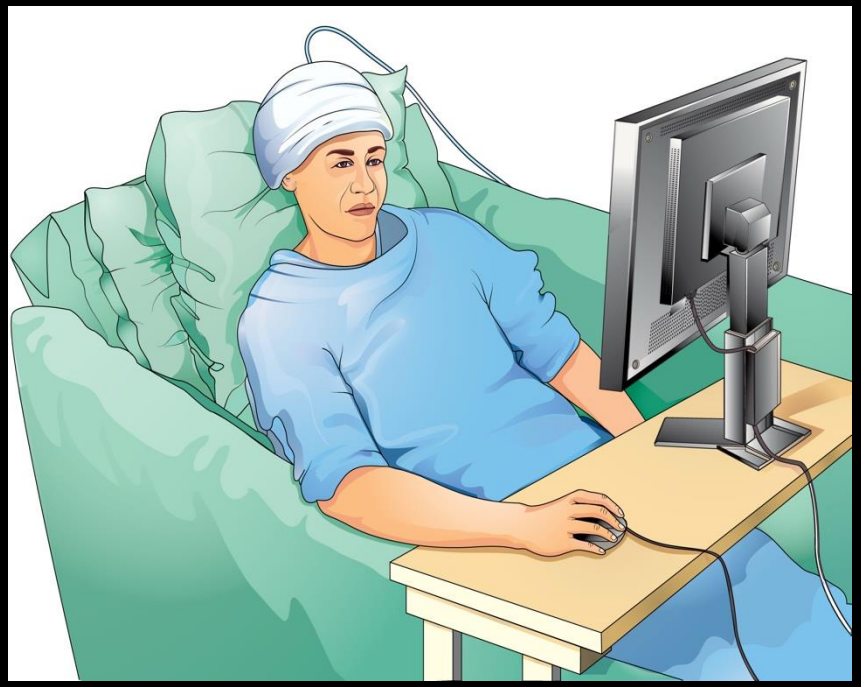
time (seconds)

Problem: fMRI is slow and indirect



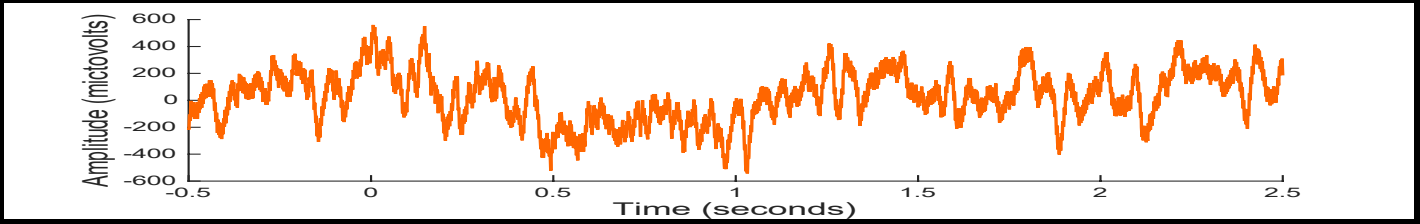
Mégevand, Mercier, Groppe, Zion Golumbic, Mesgarani, Beauchamp, Schroeder, Mehta ***Journal of Neuroscience*** 2020.

Direct recording of brain activity from neurosurgical patients with epilepsy



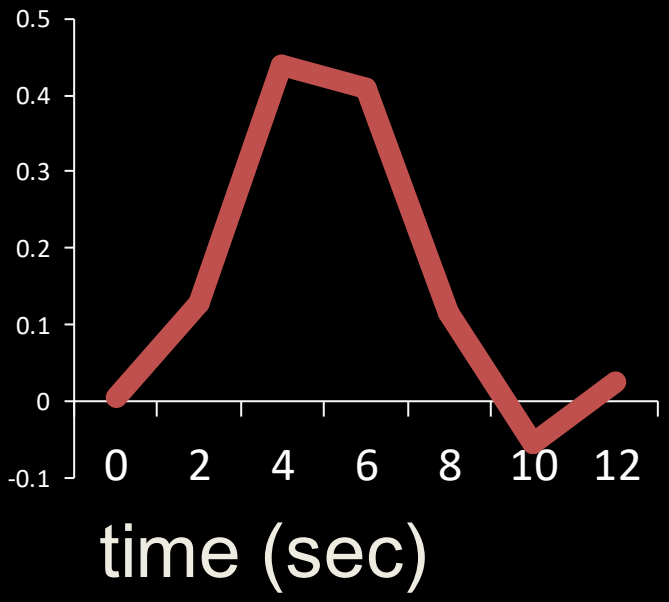
Temporal resolution

intracranial EEG: **milliseconds**



fMRI: **seconds**

multiband EPI



Neurosurgeon collaborators

Daniel Yoshor, M.D.

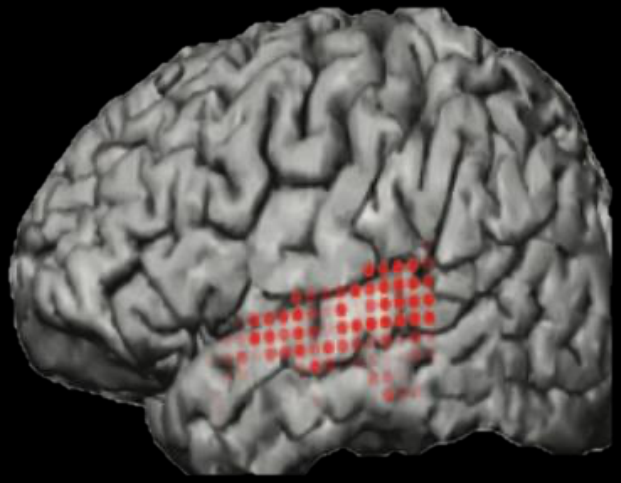
Sameer Sheth, M.D., Ph.D.

H. Isaac Chen, M.D.

intracranial EEG (iEEG) consists of two techniques

electrocorticography (ECoG)

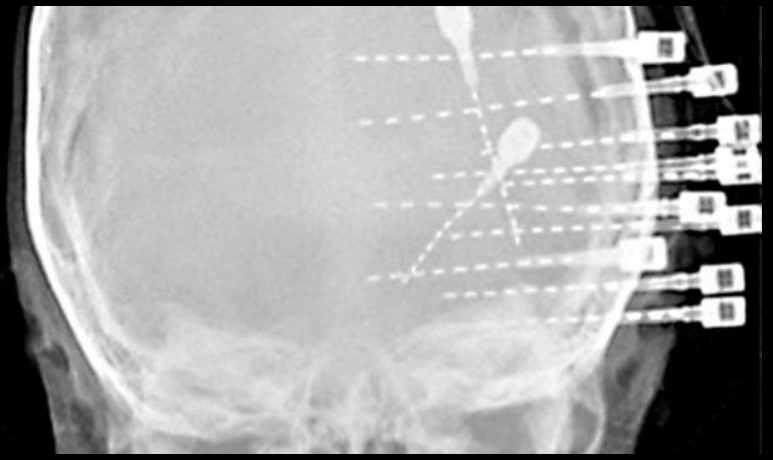
surface electrodes



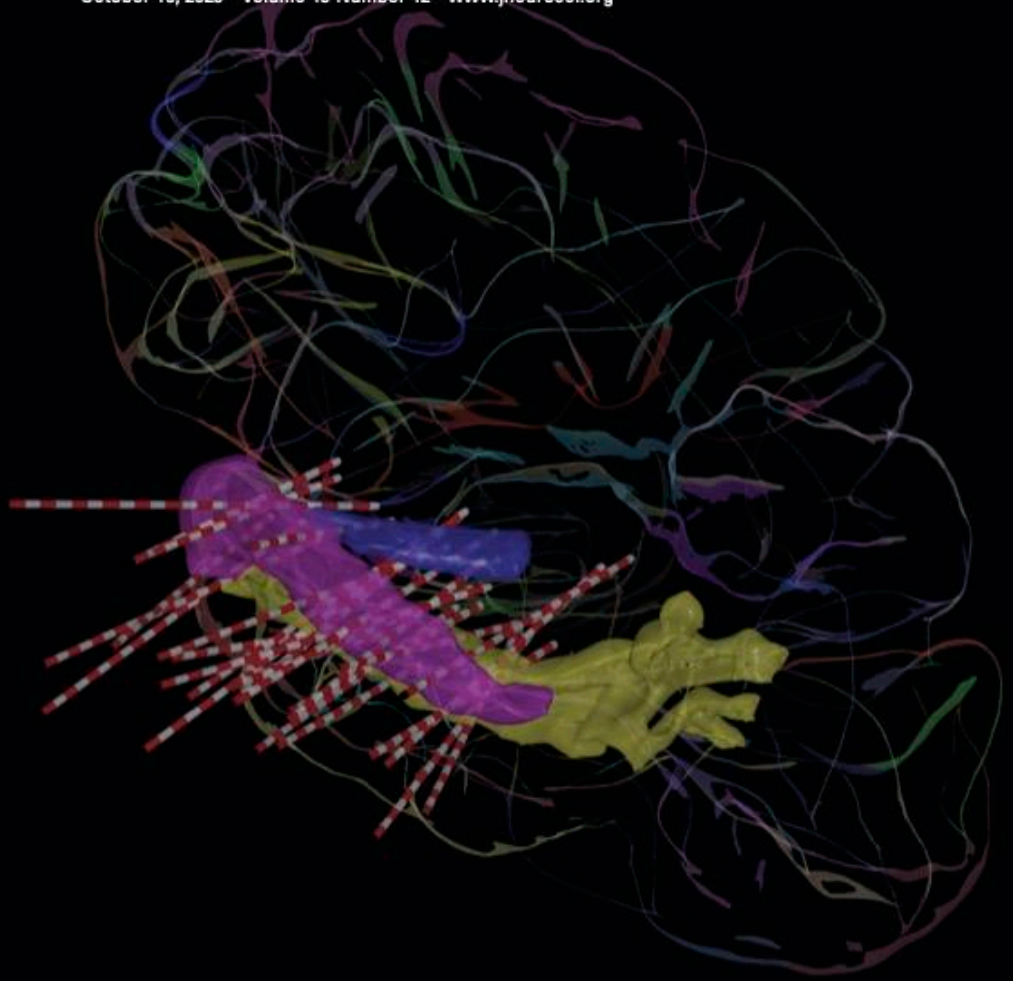
Yi, Leonard, Chang; Mesgarani *et al.;* Hamilton *et al.;* Karas *et al.*, Schepers *et al., etc.*

stereotactic EEG (sEEG)

penetrating electrodes



JNeurosci
THE JOURNAL OF NEUROSCIENCE
October 15, 2025 • Volume 45 Number 42 • www.jneurosci.org



SfN SOCIETY for NEUROSCIENCE

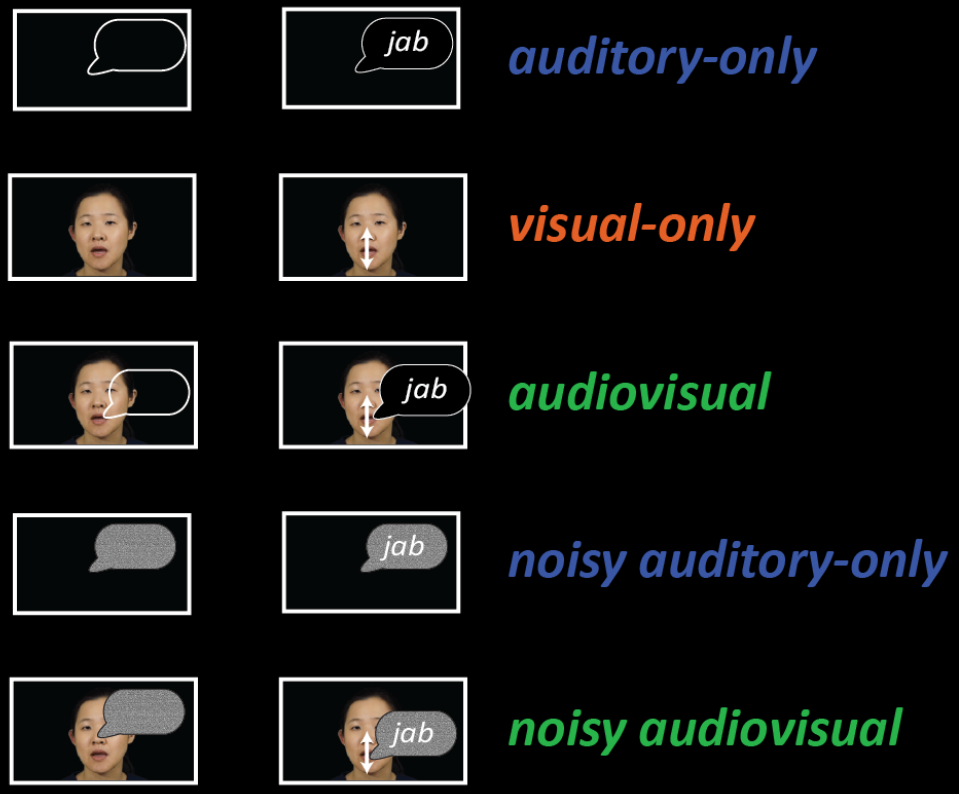
Stereoelectroencephalography Reveals Neural Signatures of Multisensory Integration in the Human Superior Temporal Sulcus during Audiovisual Speech Perception

Yue Zhang,[1,2*] John F. Magnotti,[1*] Xiang Zhang,[1] Zhengjia Wang,[1] Yingjia Yu,[1] Kathryn A. Davis,[3] Sameer A. Sheth,[2] H. Isaac Chen,[1] Daniel Yoshor,[1] and Michael S. Beauchamp[1]

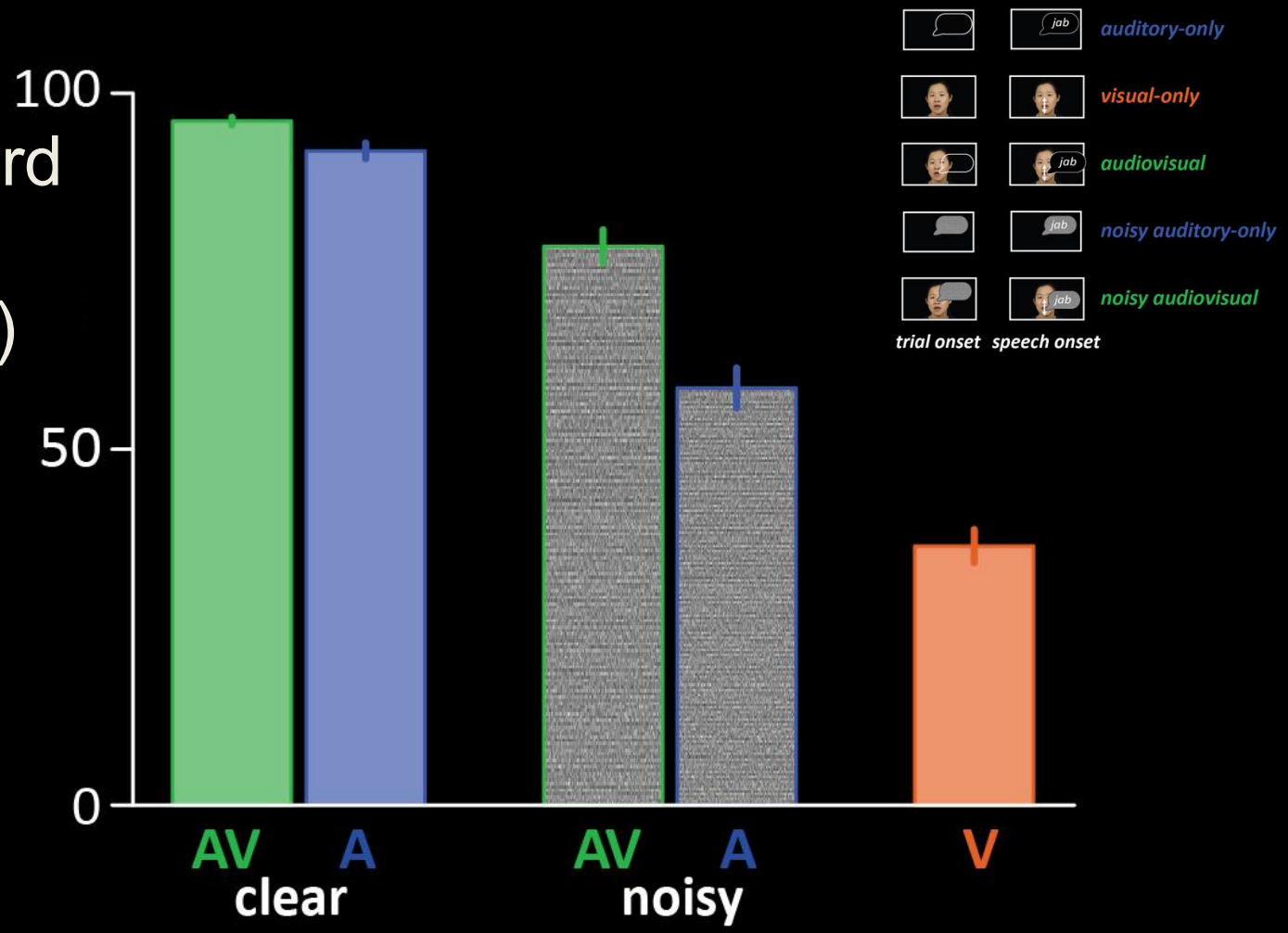
Yue Zhang

Experimental design

- n = 42 epilepsy patients
- single word stimuli

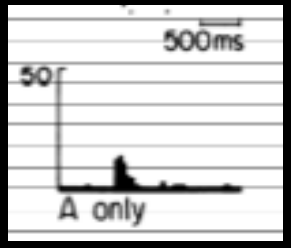


Whole-Word Accuracy (% correct)

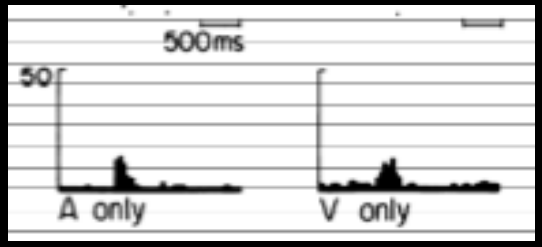


Neural signatures of multisensory integration

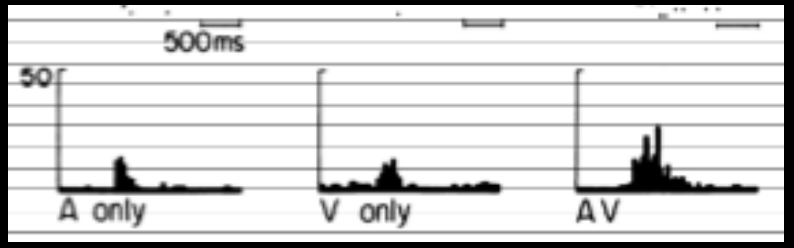
- unisensory response to auditory (A)



- unisensory response to visual (V)



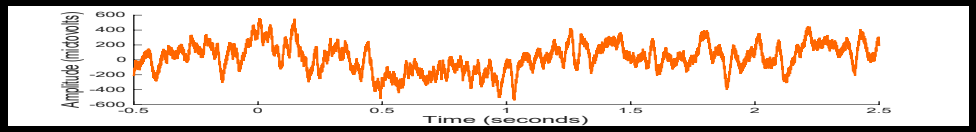
- larger multisensory response (AV)



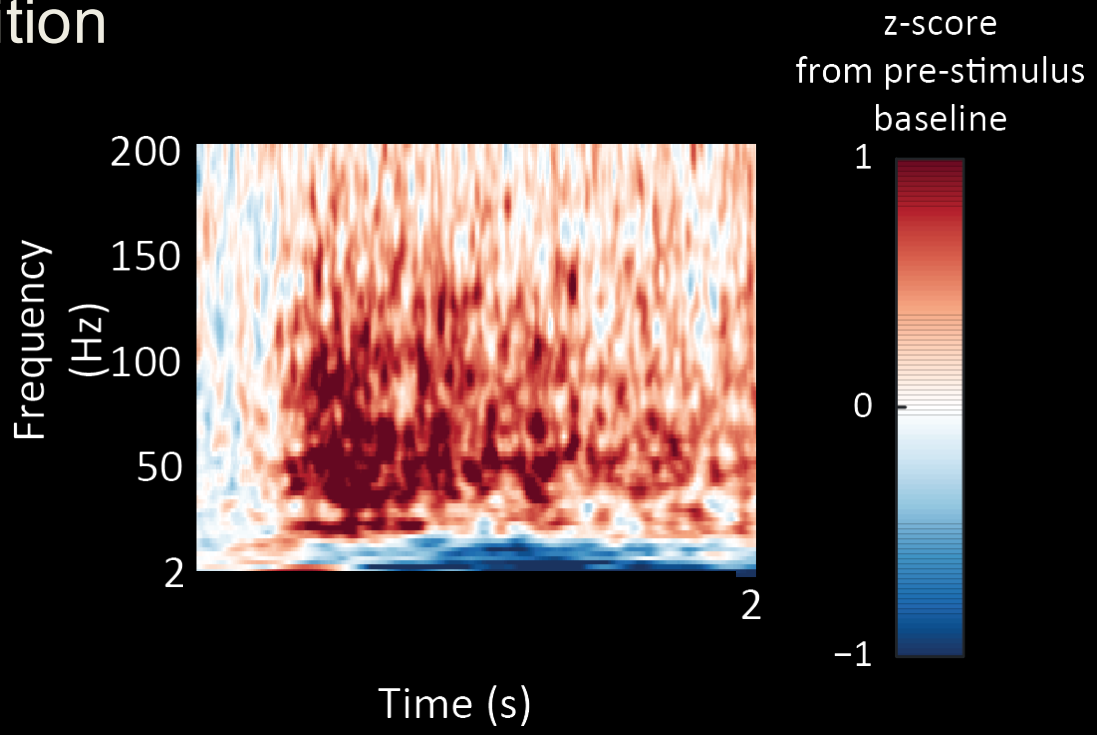
Meredith, A.M., and Stein, B.E. Journal of Neurophysiology (1986).

iEEG signal processing

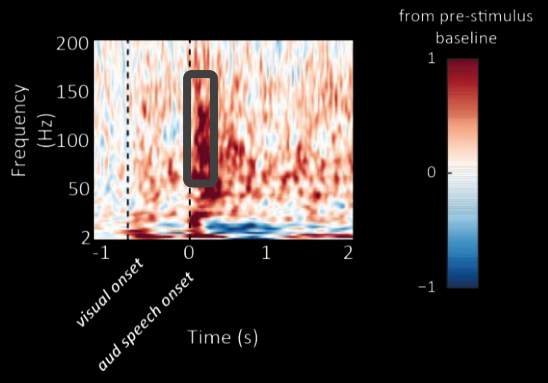
Raw Voltage Signal

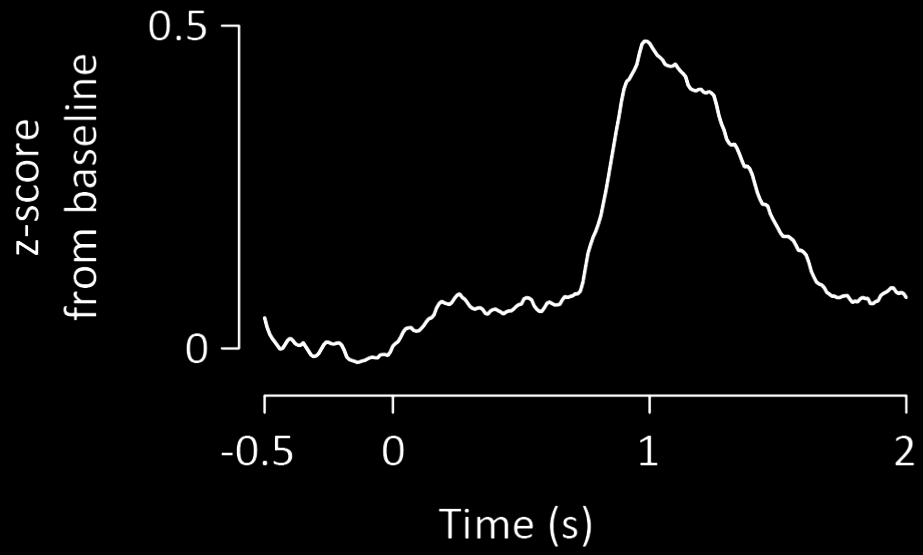


Spectral Decomposition



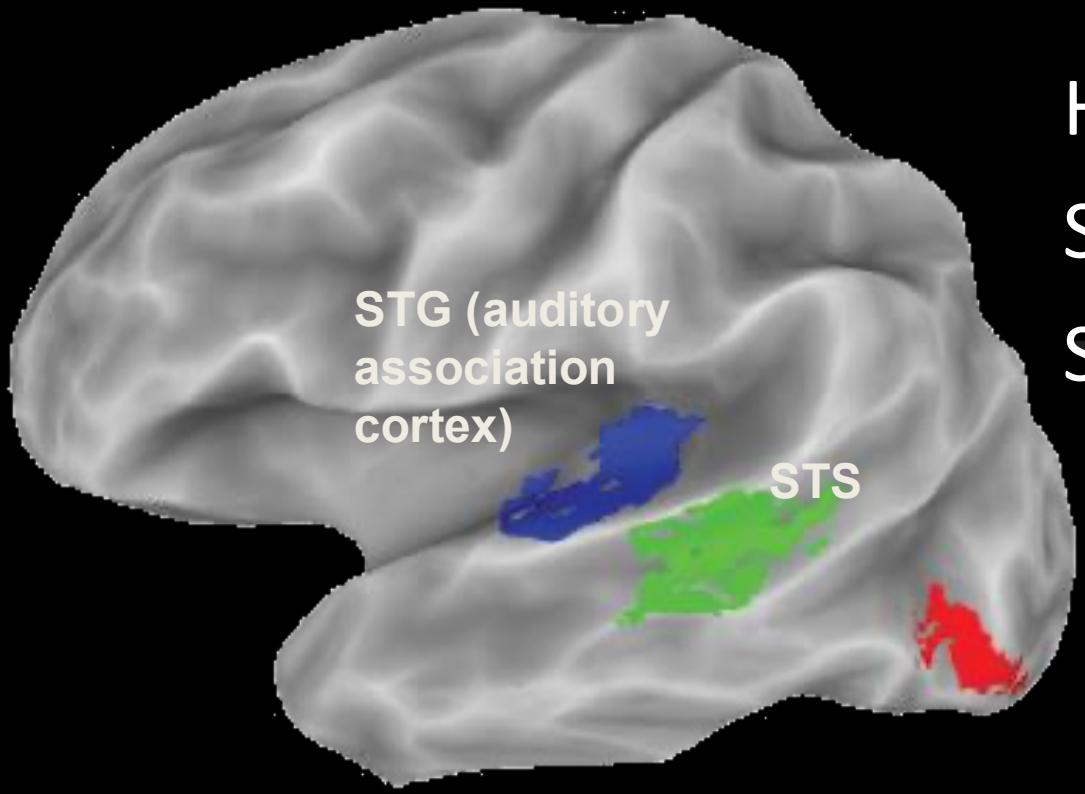
analysis: broadband high-frequency amplitude (BHA)





BHA ~ MUA *e.g.* Ray and Maunsell *PLOS Biology (*2011)

STG and STS



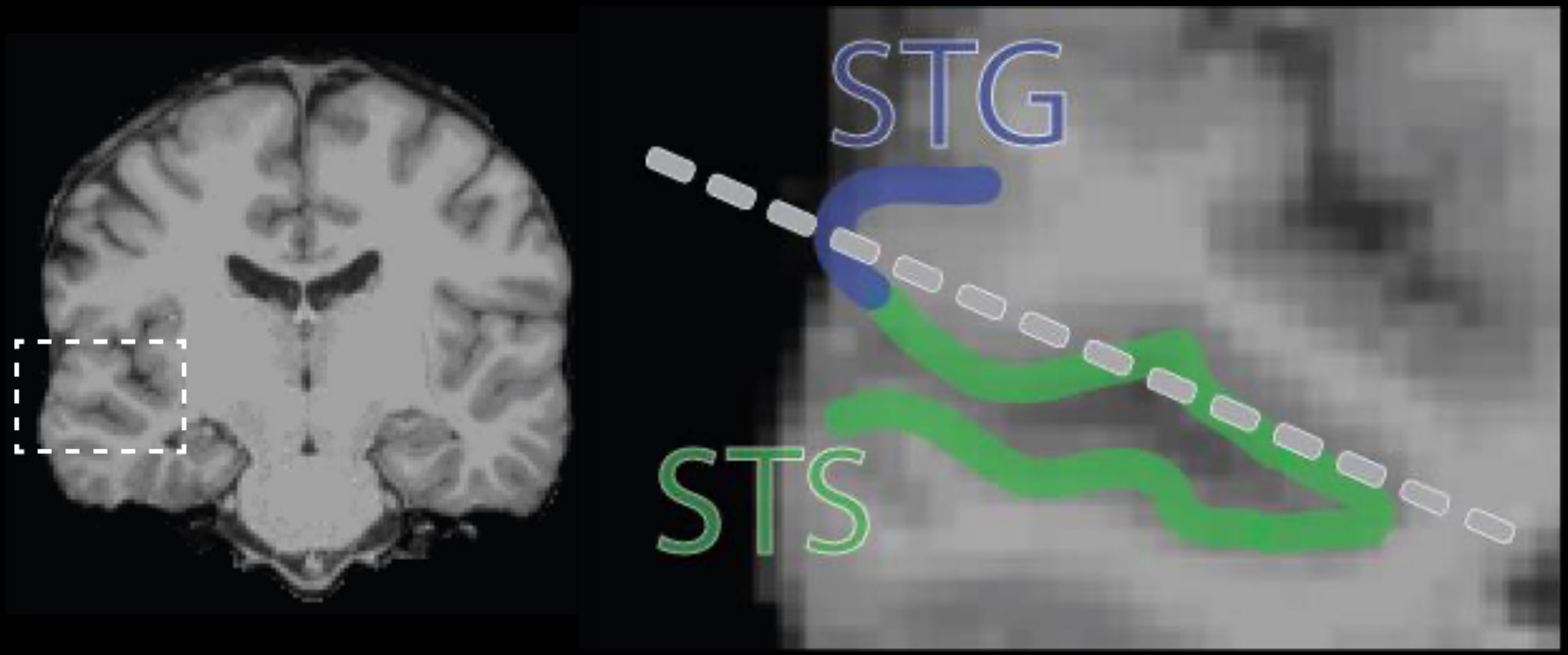
Hypothesis:

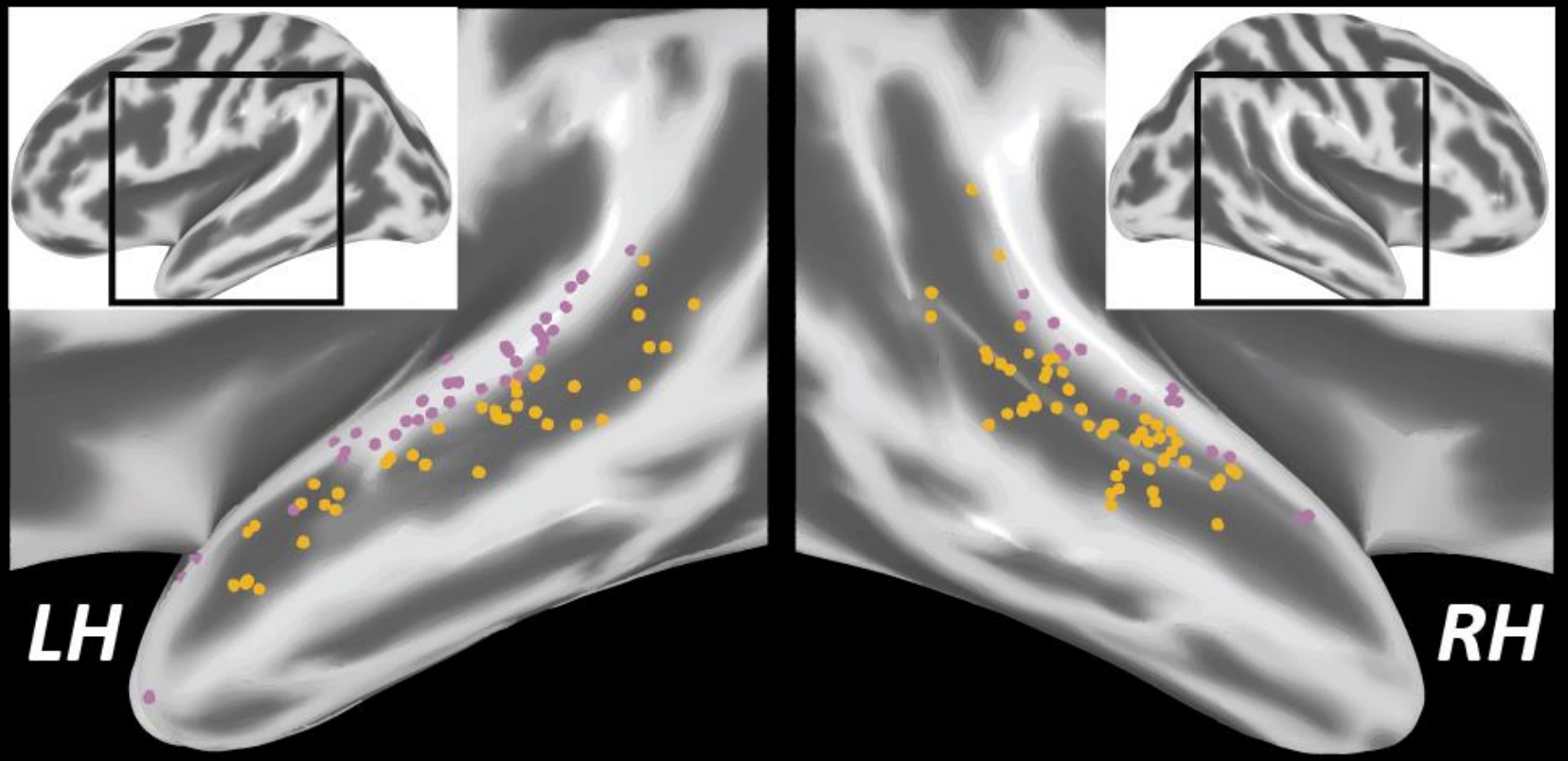
STG: auditory

STS: multisensory

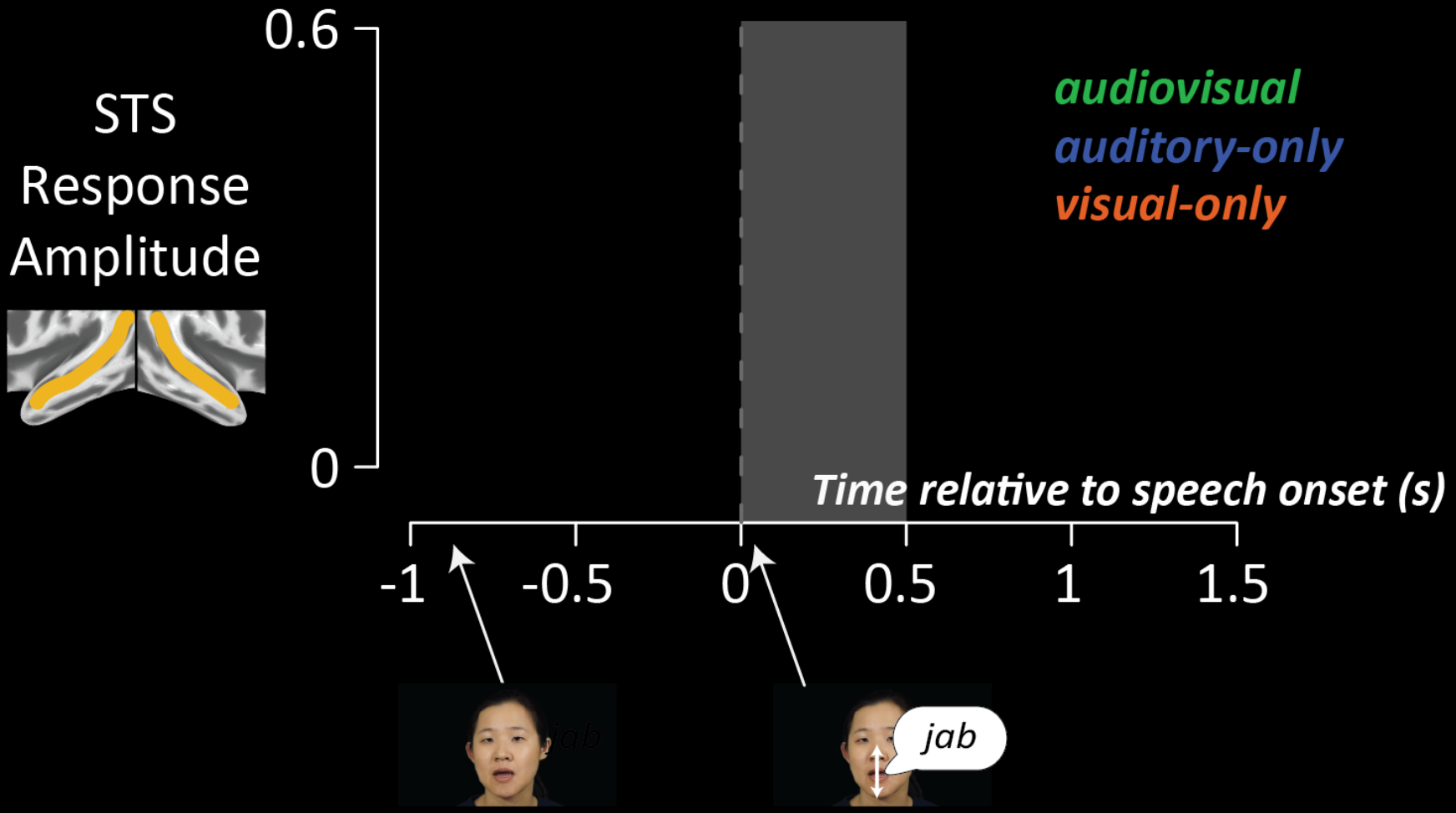
Beauchamp, *Springer Handbook of Auditory Research (2019)*

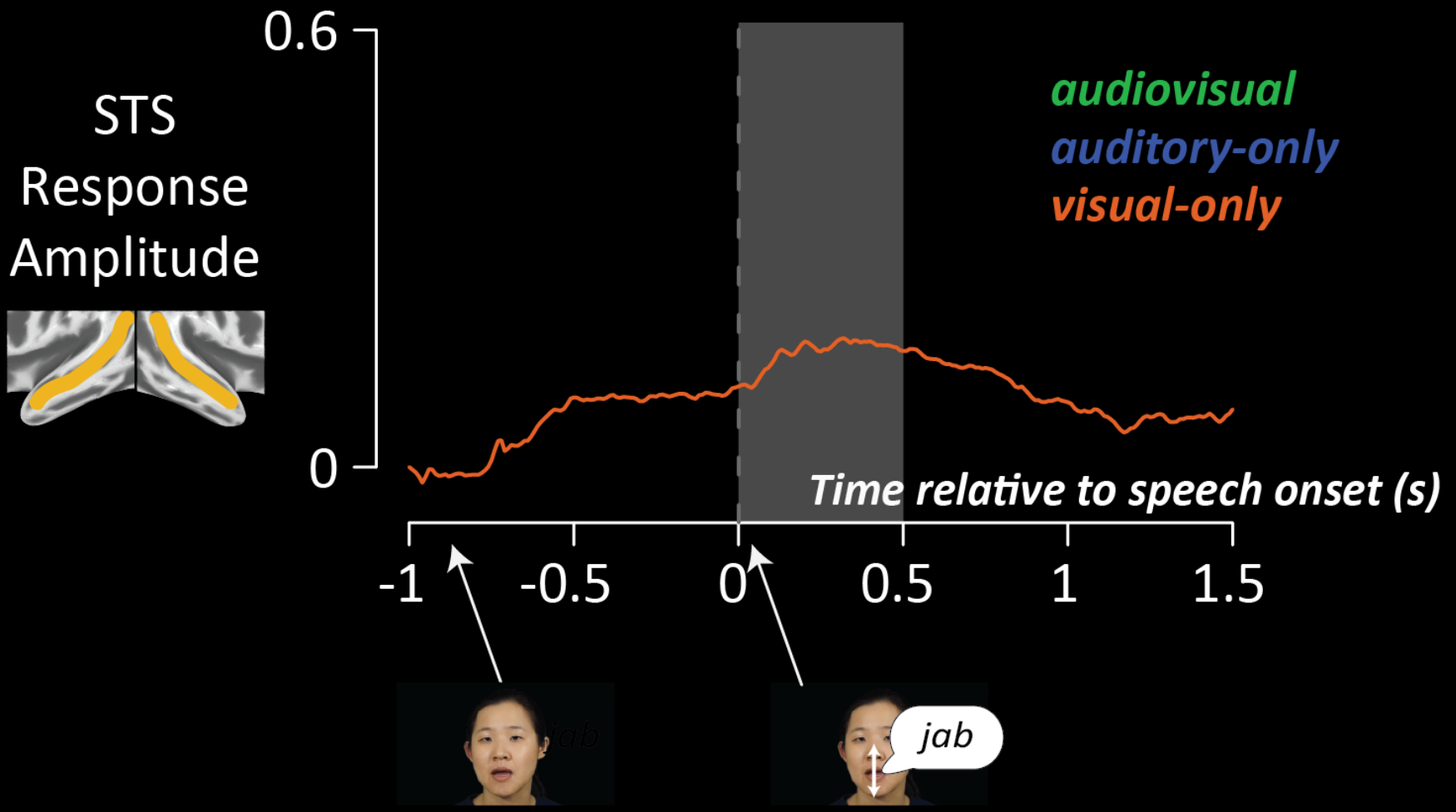
STG and STS

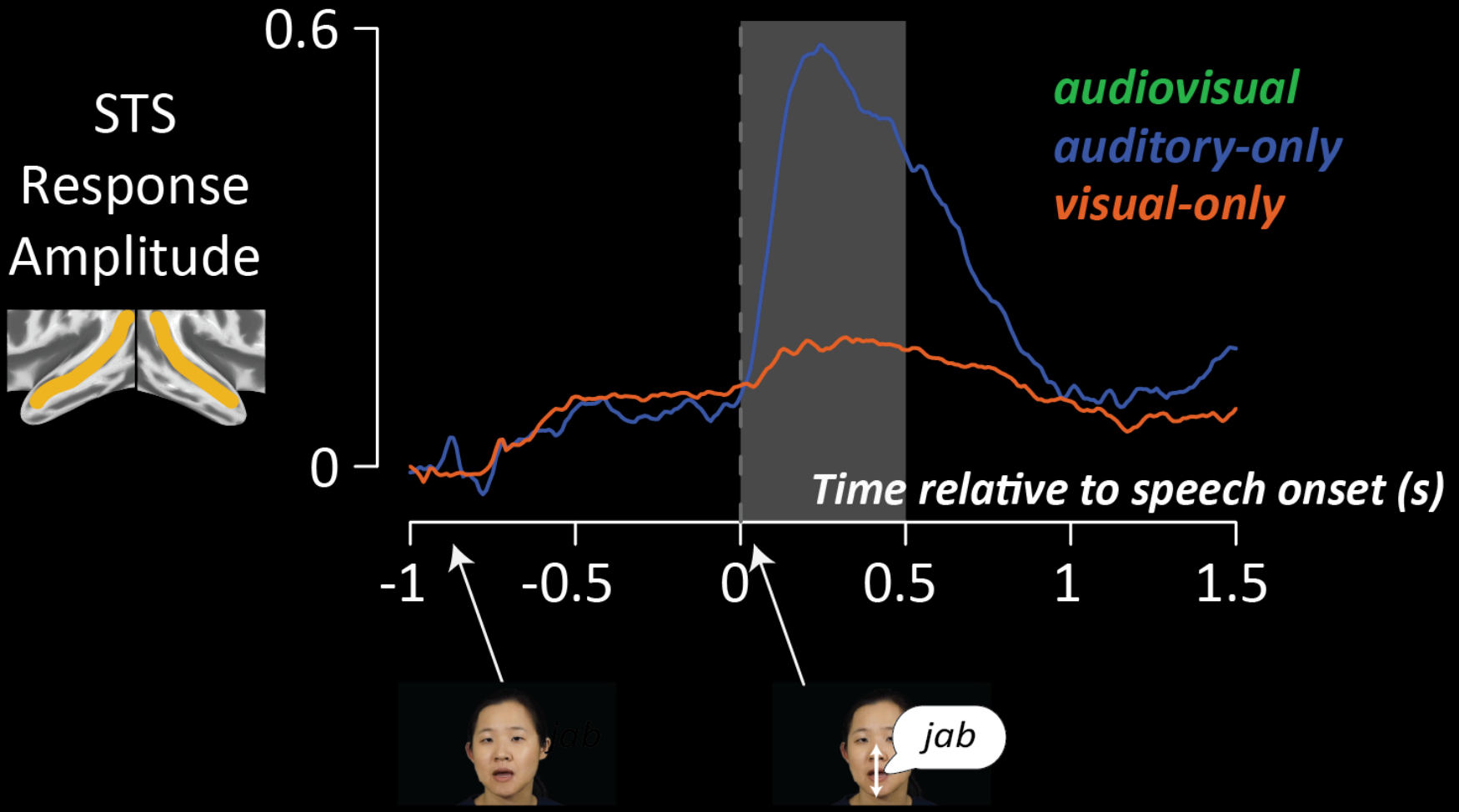


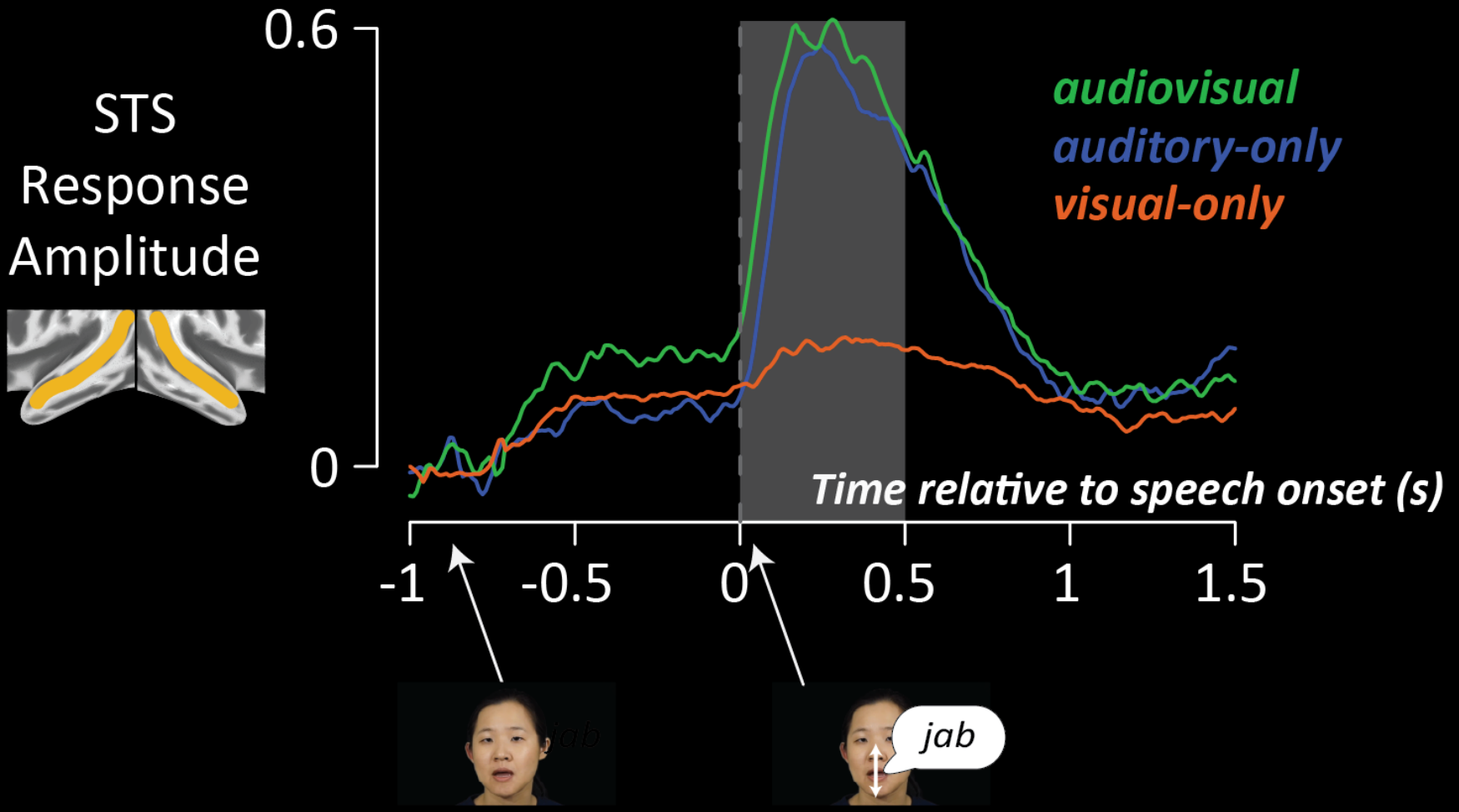


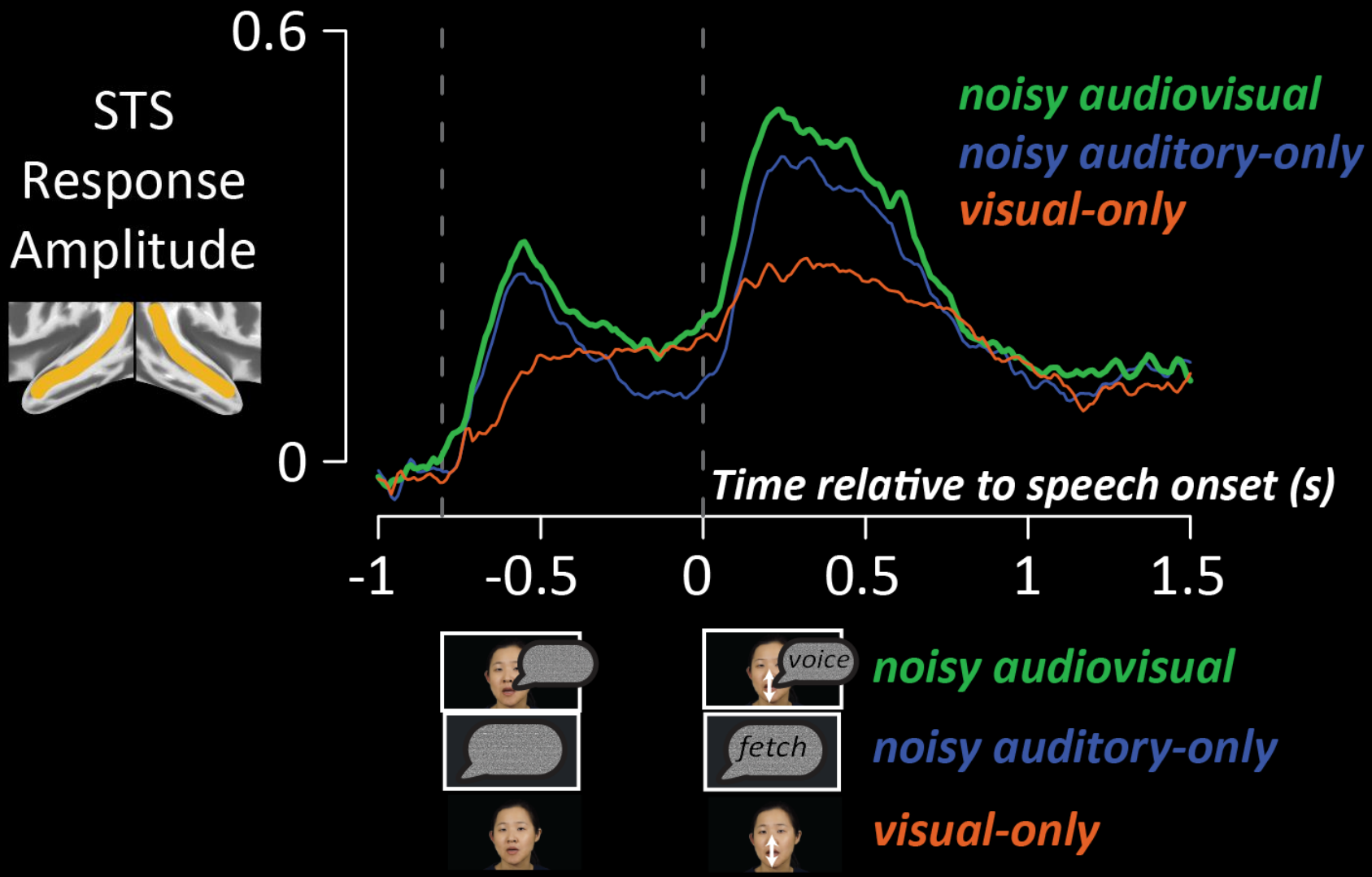
STG bimodal electrodes (n = 53)
STS bimodal electrodes (n = 93)


STS
Response
Amplitude
0.6
0
audiovisual
auditory-only
visual-only
Time relative to speech onset (s)
-1
-0.5
0
0.5
1
1.5
jab

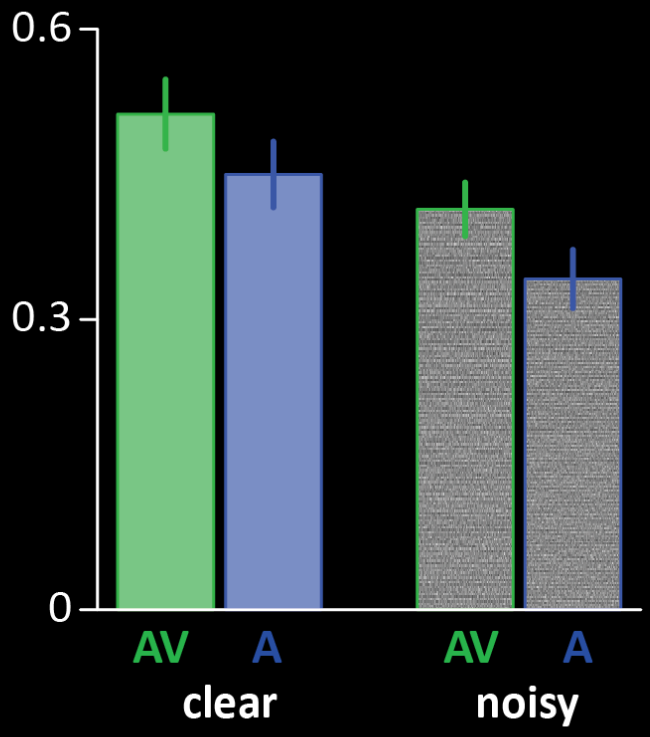

STS
Response
Amplitude
0.6
0
audiovisual
auditory-only
visual-only
Time relative to speech onset (s)
-1
-0.5
0
0.5
1
1.5
jab

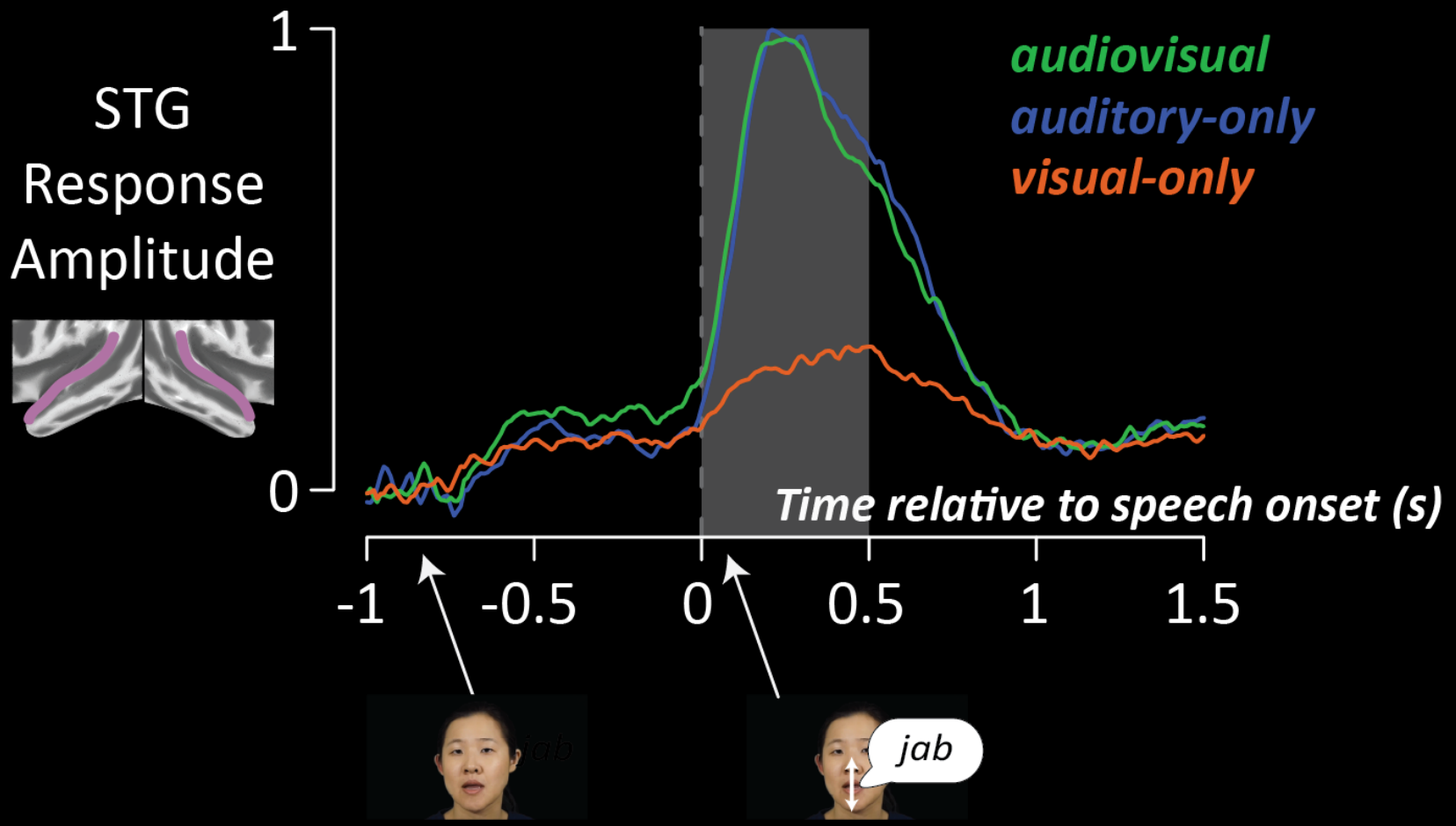

STS
Response
Amplitude
0.6
0
audiovisual
auditory-only
visual-only
Time relative to speech onset (s)
-1
-0.5
0
0.5
1
1.5
jab


STS
Response
Amplitude
0.6
0
audiovisual
auditory-only
visual-only
Time relative to speech onset (s)
-1
-0.5
0
0.5
1
1.5
jab

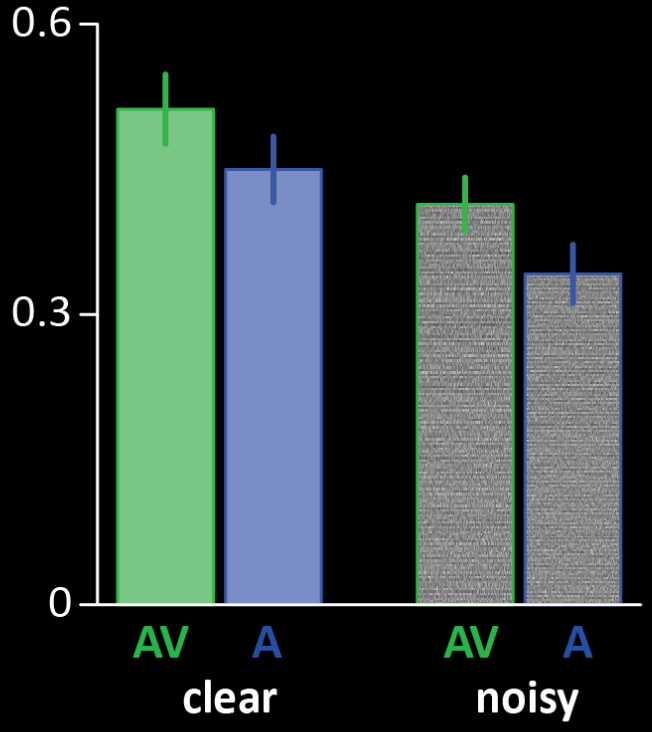

STS
Response
Amplitude
0.6
0
noisy audiovisual
noisy auditory-only
visual-only
Time relative to speech onset (s)
-1
-0.5
0
0.5
1
1.5
voice
fetch
noisy audiovisual
noisy auditory-only
visual-only

STS Response Amplitude

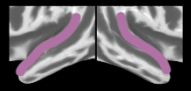


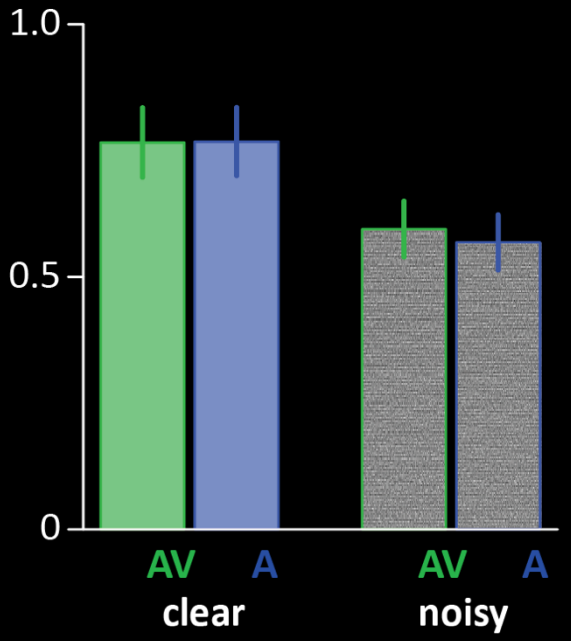

STG
Response
Amplitude
1
0
audiovisual
auditory-only
visual-only
Time relative to speech onset (s)
-1
-0.5
0
0.5
1
1.5
jab

STS Response Amplitude

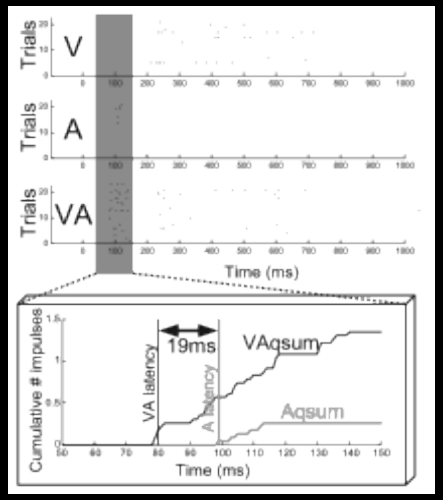


STG Response Amplitude





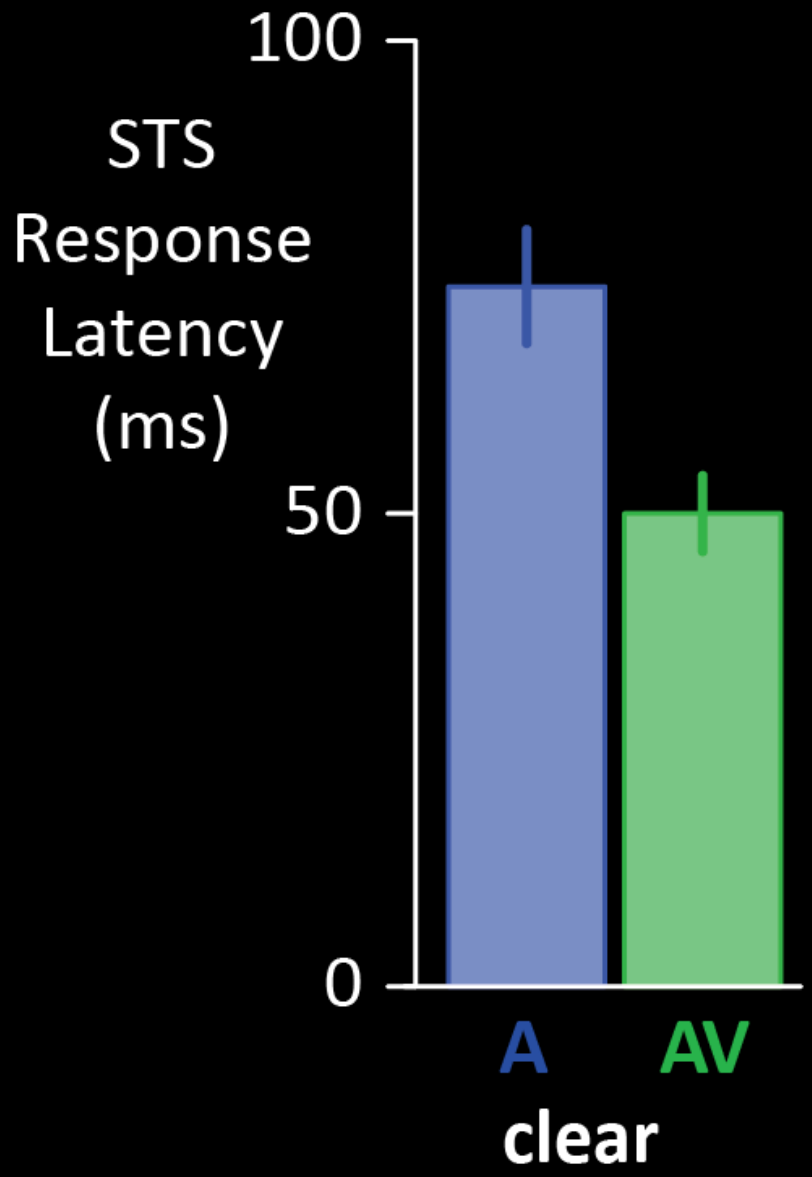
Neural signatures of audiovisual integration: #2: ***Faster*** audiovisual response

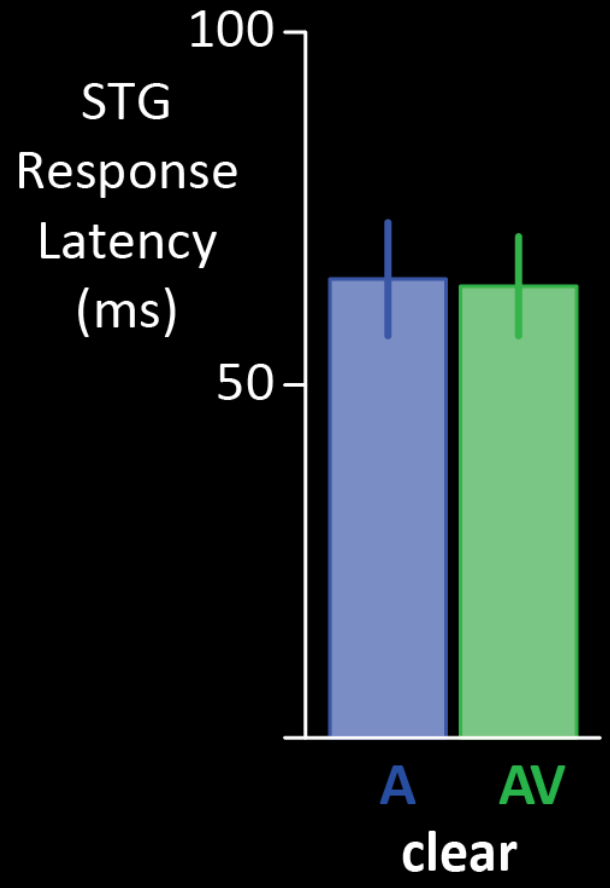


Rowland, B.A., Quessy, S., Stanford, T.R., and Stein, B.E. Journal of Neuroscience (2007).

shorter latency

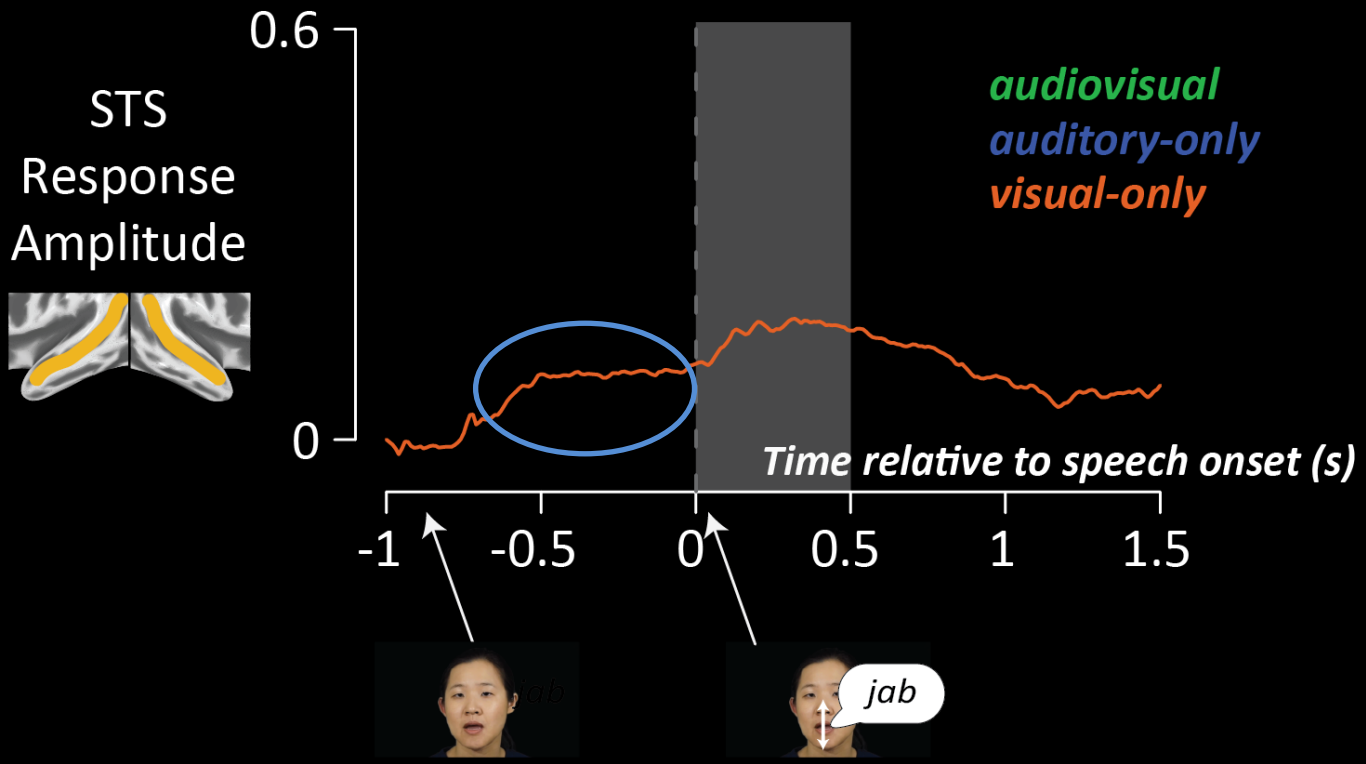
(faster) audiovisual response


100
STS
Response
Latency
(ms)
50
0
A
AV
clear


100
STG
Response
Latency
(ms)
50
A
AV
clear

But the visual system is slow, how can it speed auditory speech processing?

Answer: preparatory gestures



Chandrasekaran *et al., PLOS Computational Biology* (2009)
but see Schwartz and Savariaux *PLOS Computational Biology* (2014)

Section summary

- sEEG provides evidence for multisensory integration of speech in STS but not STG
  - unisensory responses to auditory and visual speech
  - Greater response amplitude for audiovisual speech
  - Shorter response latency for audiovisual speech

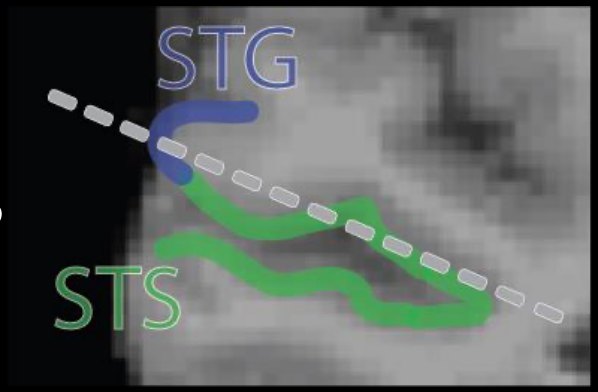
Questions?

Talk Overview

- Introduction to audiovisual speech perception
- Behavior
  - *Magnotti et al. (under review)*
- Neural substrates
  - *Zhang et al. (Journal of Neuroscience, 2025)*
- Deep neural network models
  - *Ma et al. (Psychonomic Bulletin and Review, 2026)*

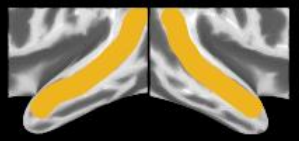
BOLD fMRI

intracranial EEG

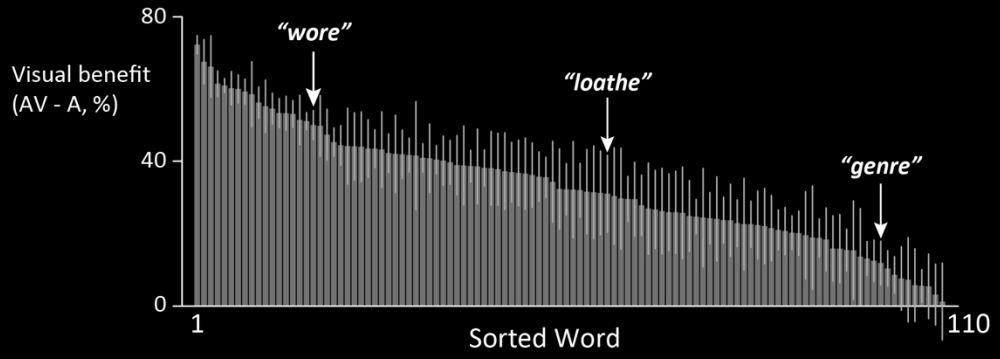


Model motivation

STS Response Amplitude



models

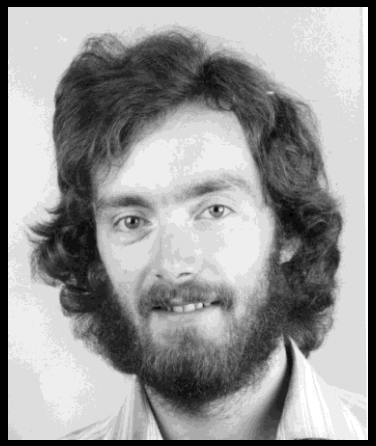


consonant in the first syllable of each utterance. E
combined three replications of its auditory-visu
pectively; errors were unsystematic. Under the
conduction deafness
's by adding masking
al power in all audible

Harry McGurk in 1976

b. 23 February 1936
d. 17 April 1998

John MacDonald in 1976



Symposium on the 40th Anniversary of the McGurk Effect (2016)

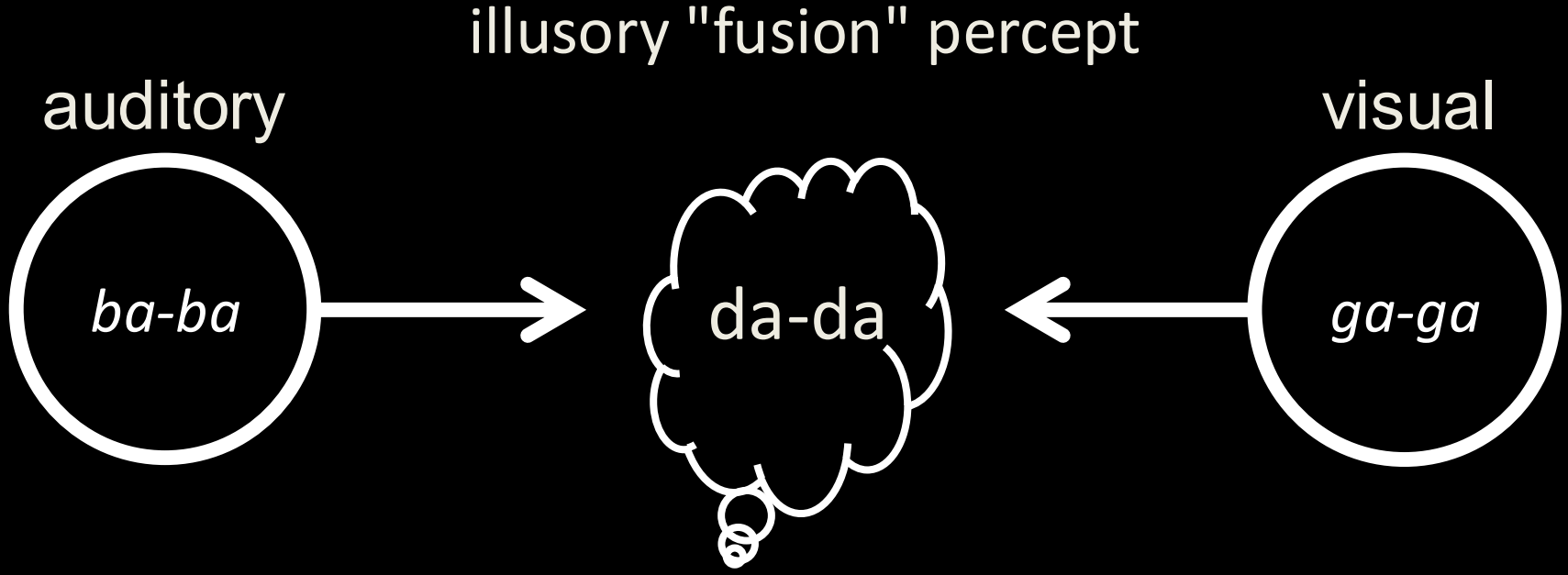
Figure 1. Speakers at the fortieth anniversary symposium held on June 16th, 2016. From left: John MacDonald, Michael Beauchamp, Julia Irwin, Salvador Soto-Faraco. Jean Vroomen was not able to attend but delivered his lecture remotely.

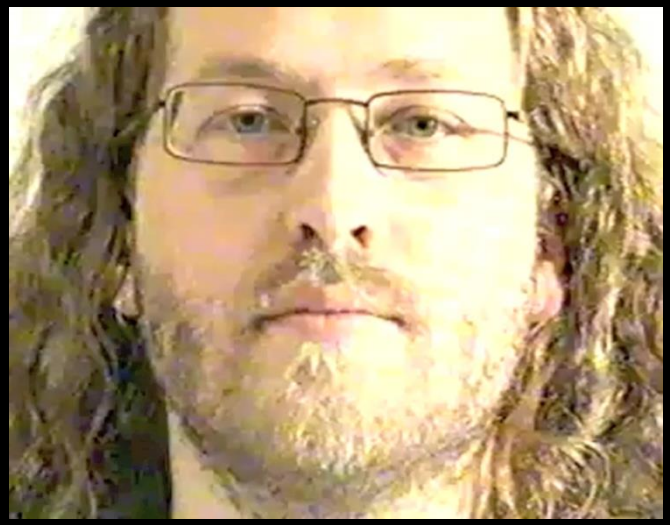
Multisensory Research 31 (2018) 1–6

Symposium on the 50th Anniversary of the McGurk Effect (2026)

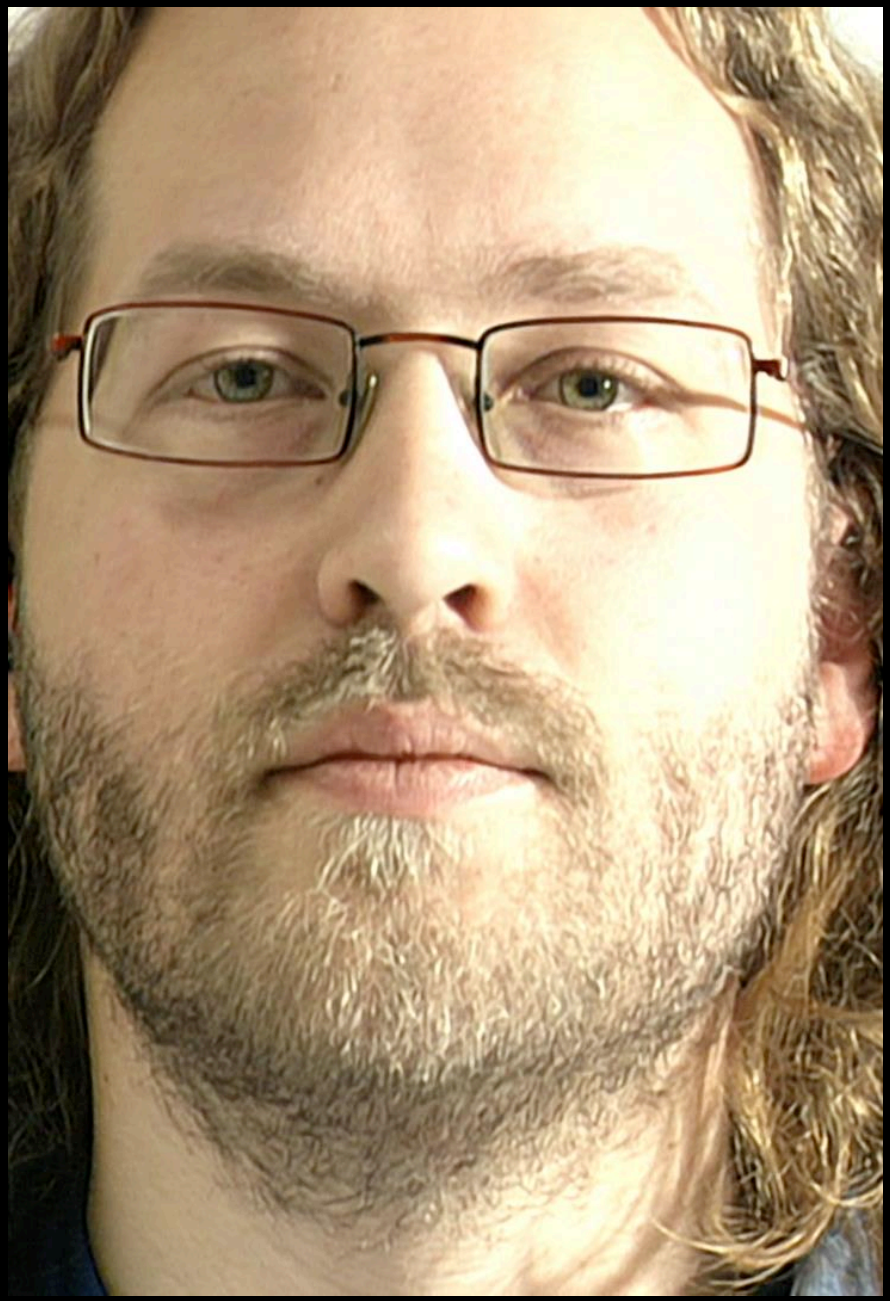
The 24th International Multisensory Research Forum will take place at the conference center **"Magazzini del Cotone"** in **Genoa, Italy** on **June 24th-27th, 2026.**

McGurk doing McGurk





Courtesy of Arnt Maasø , University of Oslo



Models

- Word models
- Bayesian causal inference models
- Deep Neural Network models (DNNs)

auditory | illusory percept | visual

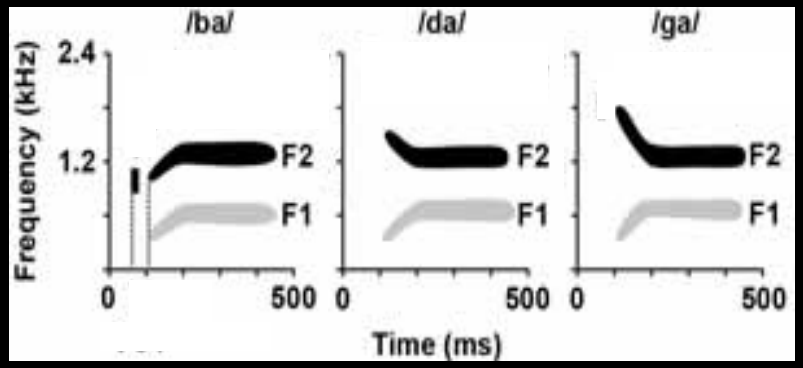
ba-ba → da-da ← *ga-ga*

Visual features *e.g.* mouth aperture

da-da ~ ga-ga

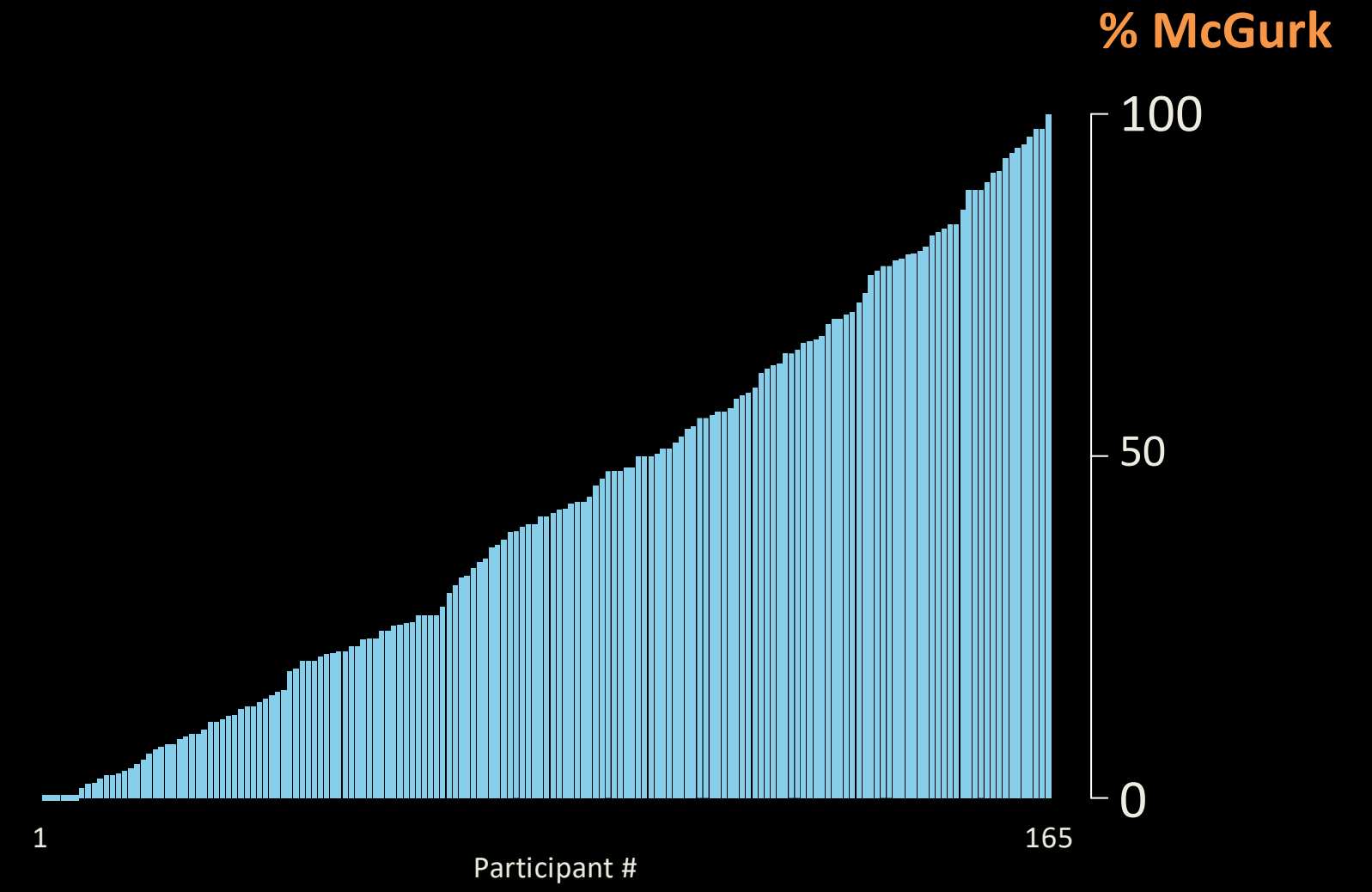
ba-ba ~ **da-da**

Acoustic features *e.g.* F2 Transition

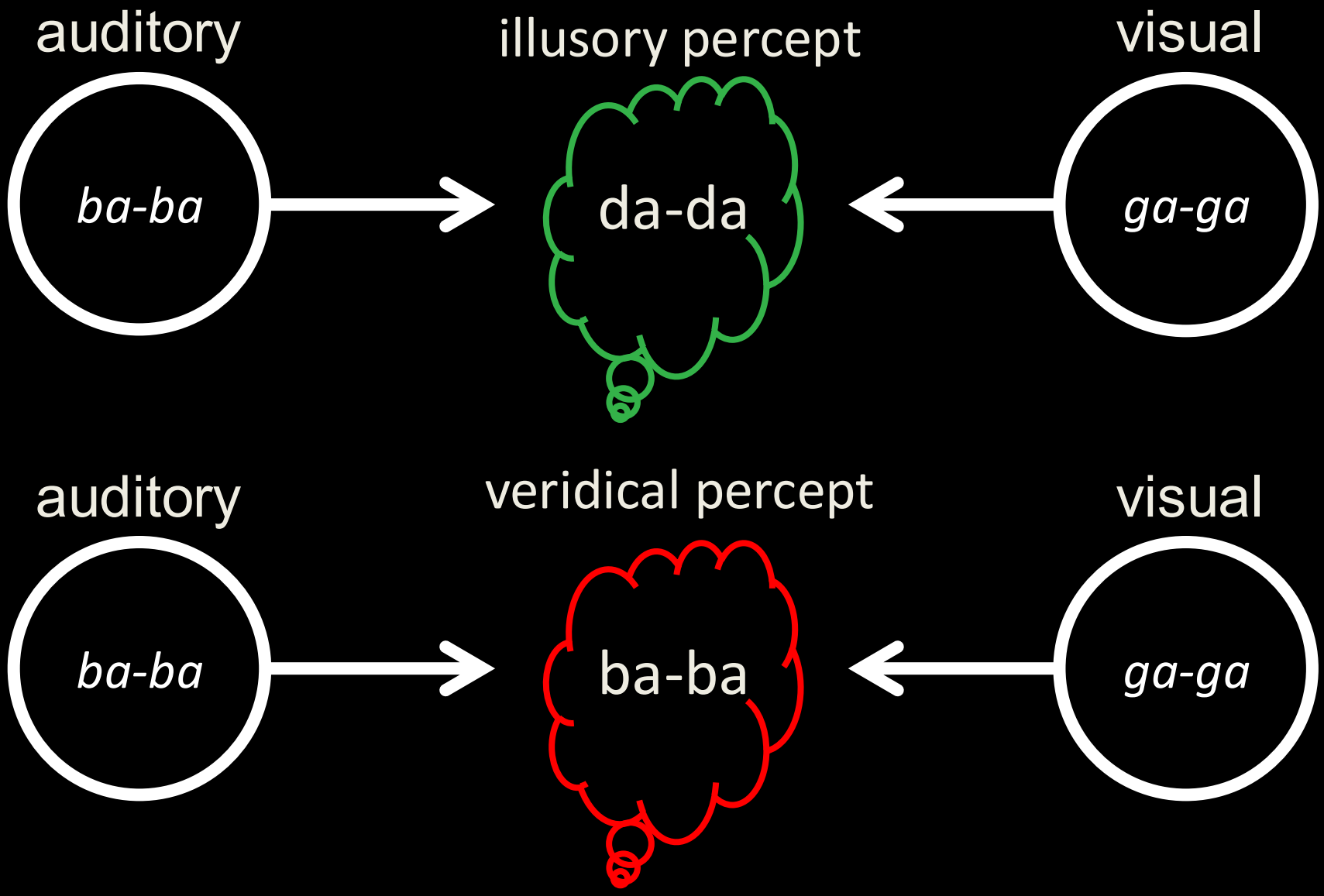


Grant PTRS-B, 2006

But: very high intersubject variability
n = 165 participants tested in person

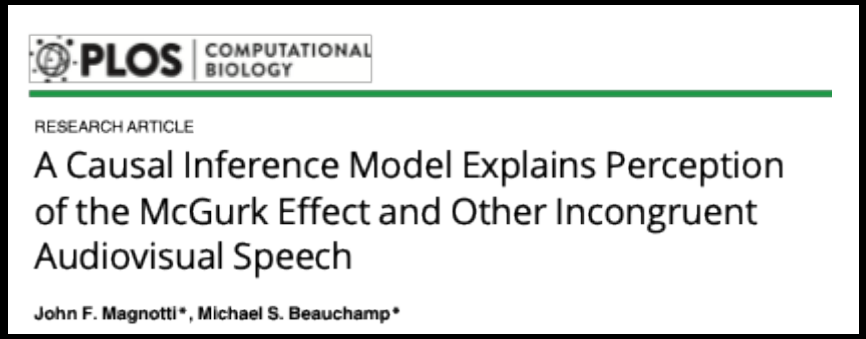


Basu Mallick, Magnotti, Beauchamp, *Psychonomic Bulletin and Review (2015)*


auditory
illusory percept
visual
ba-ba
da-da
ga-ga
auditory
veridical percept
visual
ba-ba
ba-ba
ga-ga
+ many other percepts
ah-ah
ga-ga
fa-fa

Models

- Word models
- Bayesian causal inference models
- Deep Neural Network models (DNNs)


PLOS COMPUTATIONAL BIOLOGY

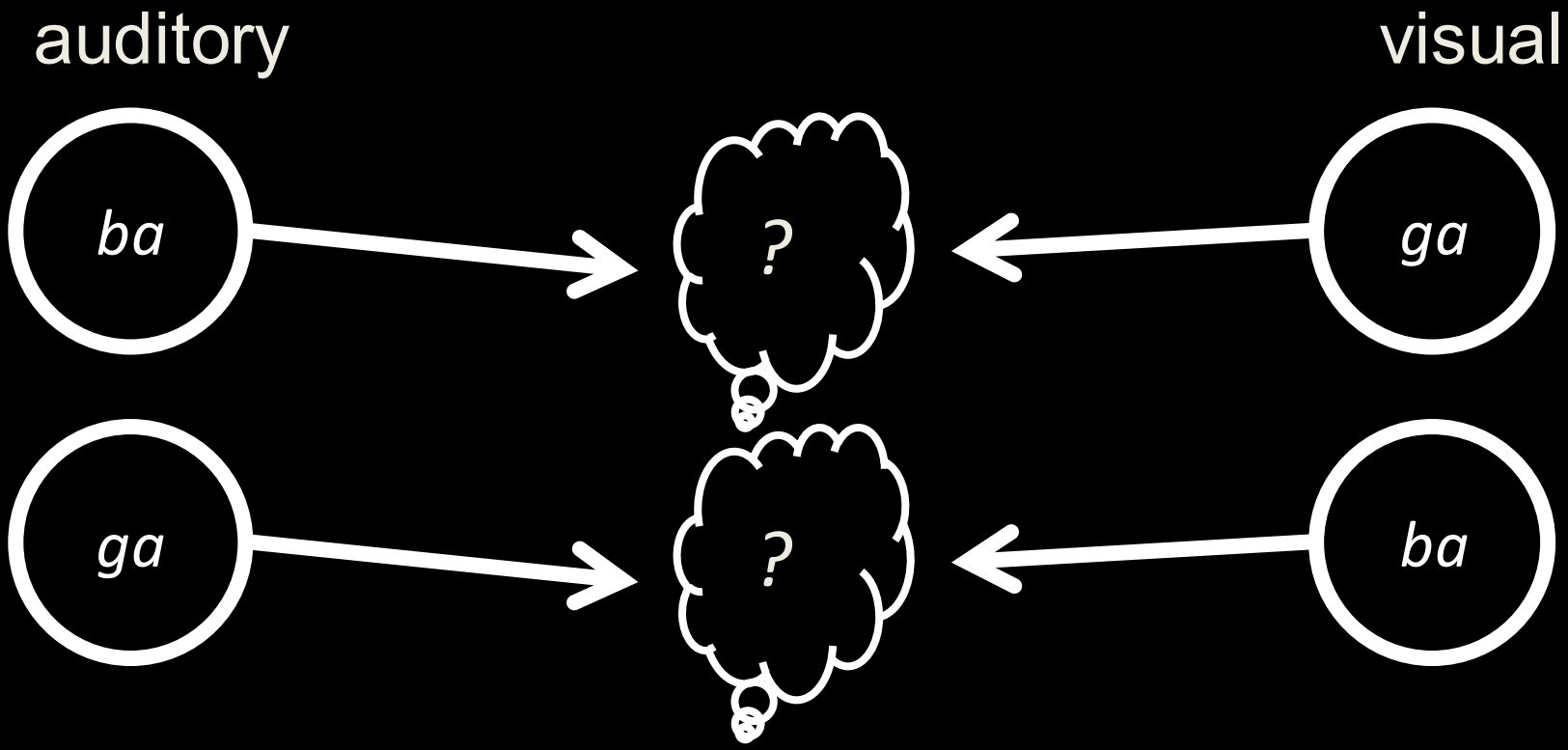
RESEARCH ARTICLE

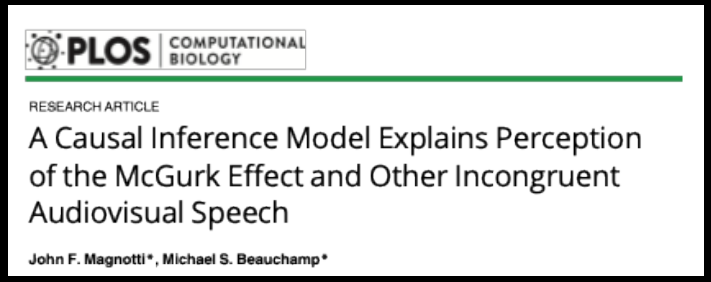
A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech

John F. Magnotti*, Michael S. Beauchamp*

Noppeney (2021)
Kording et al. (2007)
Ma et al. (2009)
Parise & Ernst (2025)
Shams & Beierholm (2025)

Bayesian models do not generalize outside of their specification



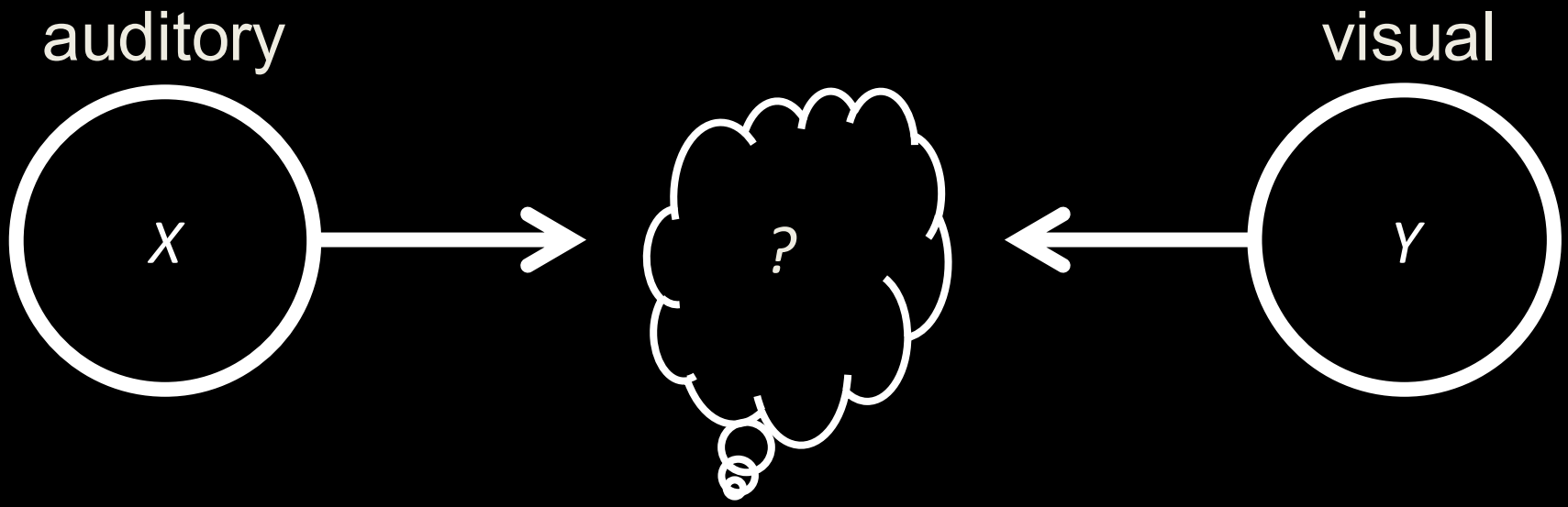

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

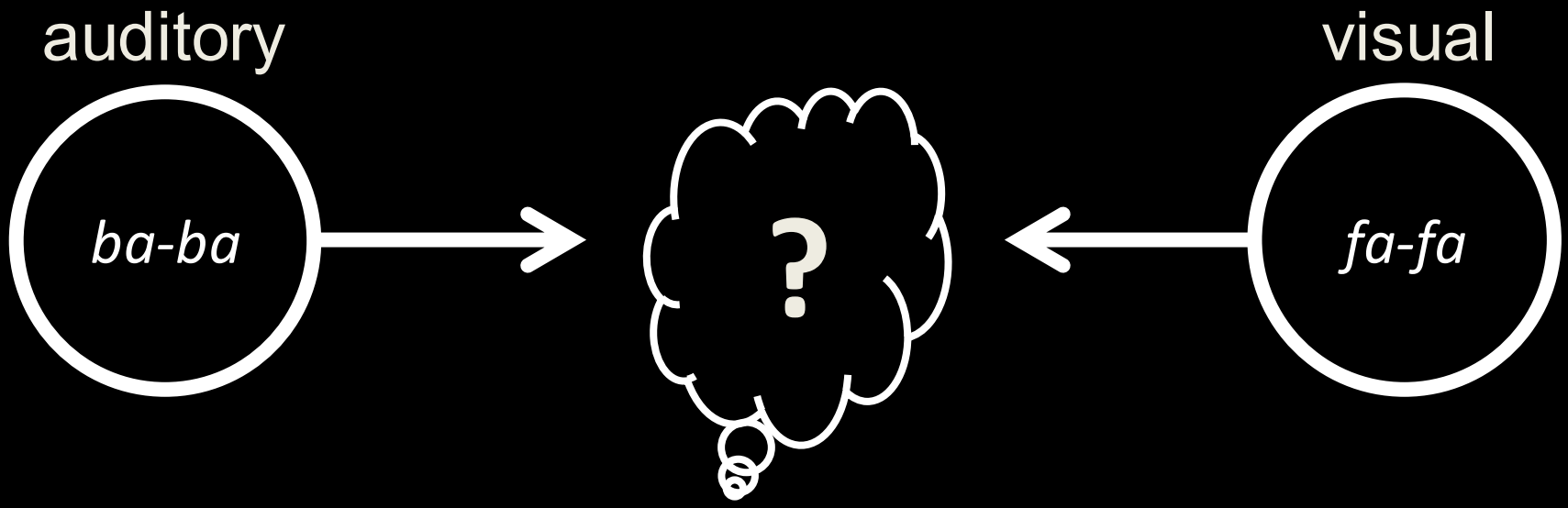
A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech

John F. Magnotti*, Michael S. Beauchamp*

Bayesian models do not generalize outside of their specification



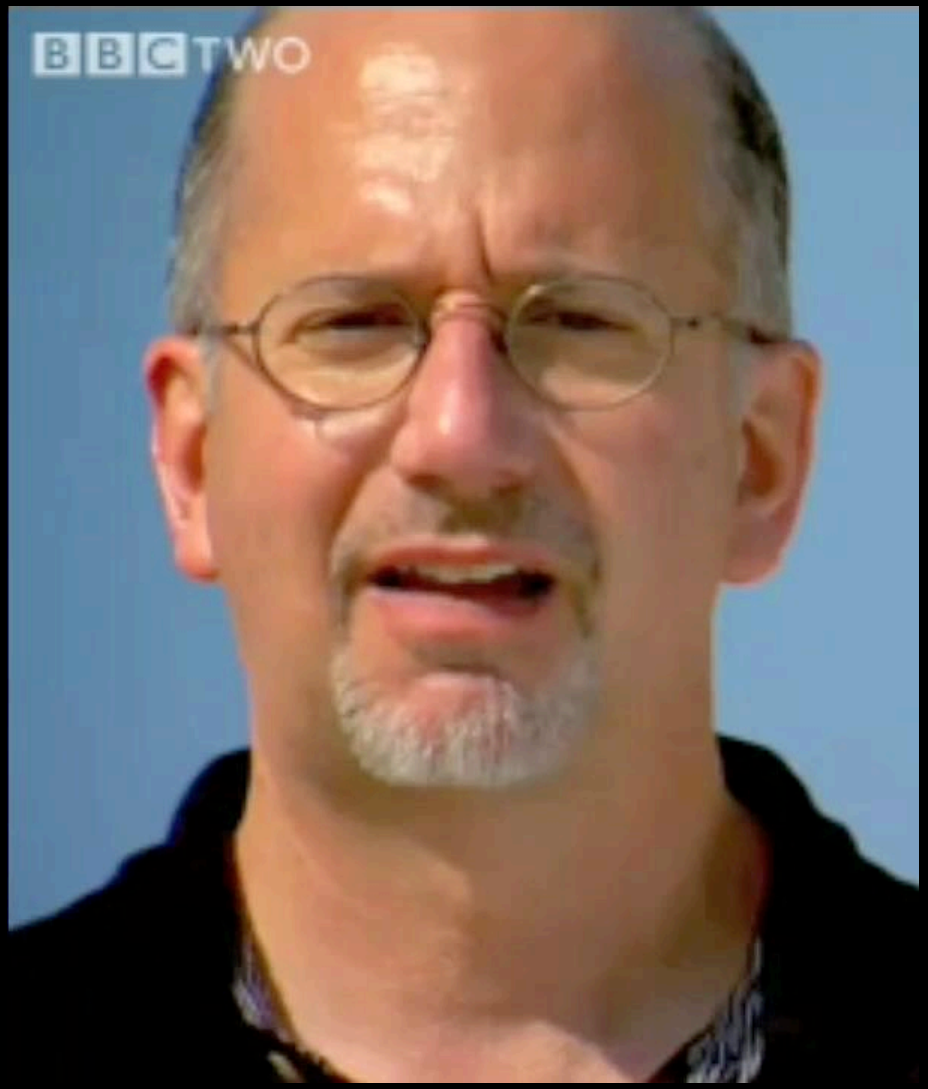
Bayesian models do not generalize outside of their specification



Auditory “ba”

Courtesy of Larry Rosenblum, UC Riverside

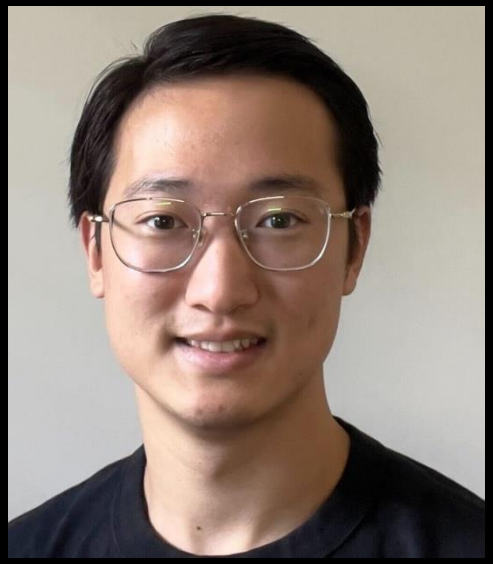
Auditory “ba” + Visual “fa”




BBC TWO

Models of audiovisual speech perception

- Word models
- Causal inference models
- Deep Neural Network models (DNNs)



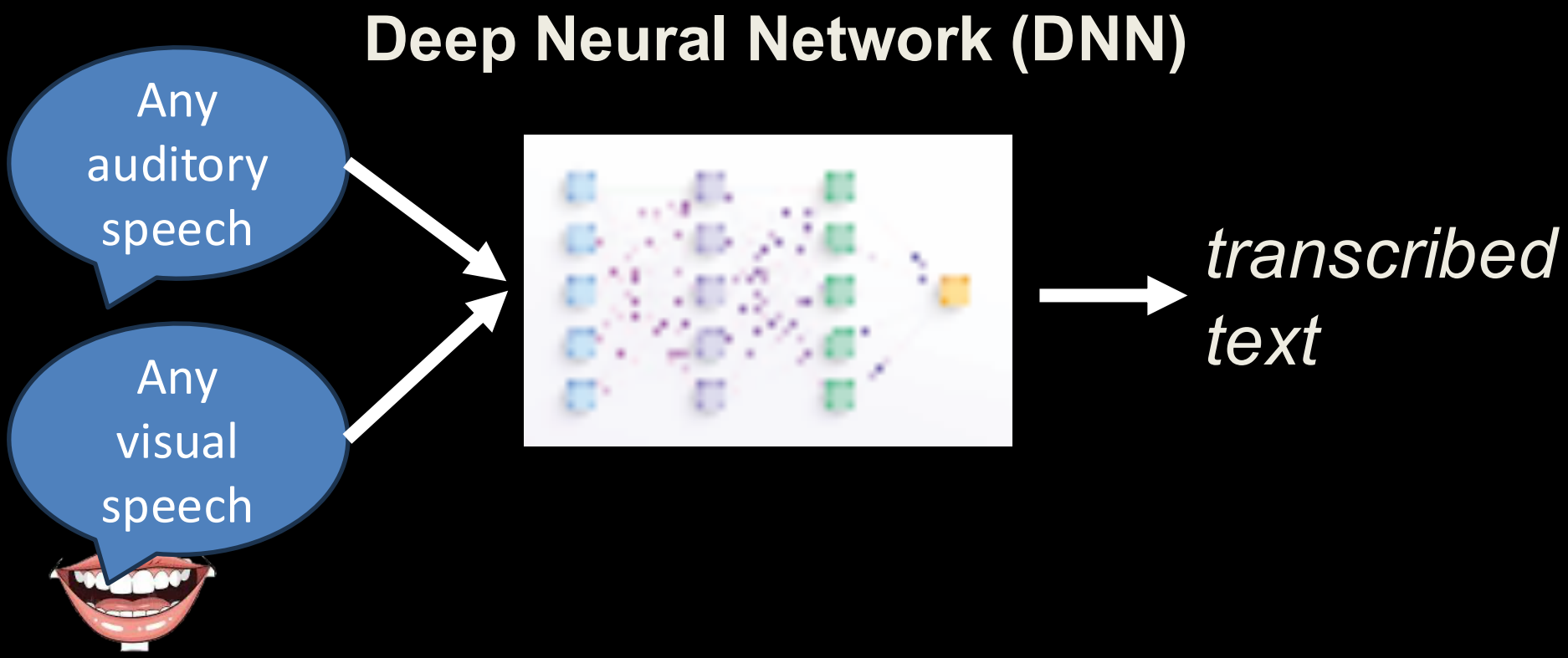
Haotian Ma

A deep neural network model of audiovisual speech recognition reports the McGurk effect

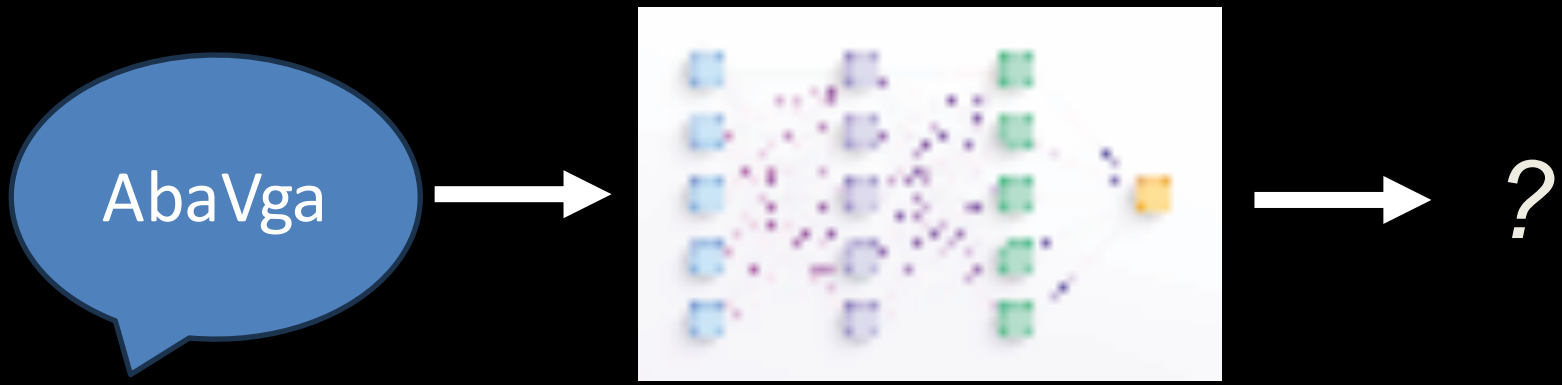
Haotian Ma[1] · Zhengjia Wang[1] · Xiang Zhang[1] · John F. Magnotti[1] · Michael S. Beauchamp[1]

Psychonomic Bulletin & Review (2026) 33:84
https://doi.org/10.3758/s13423-025-02846-8

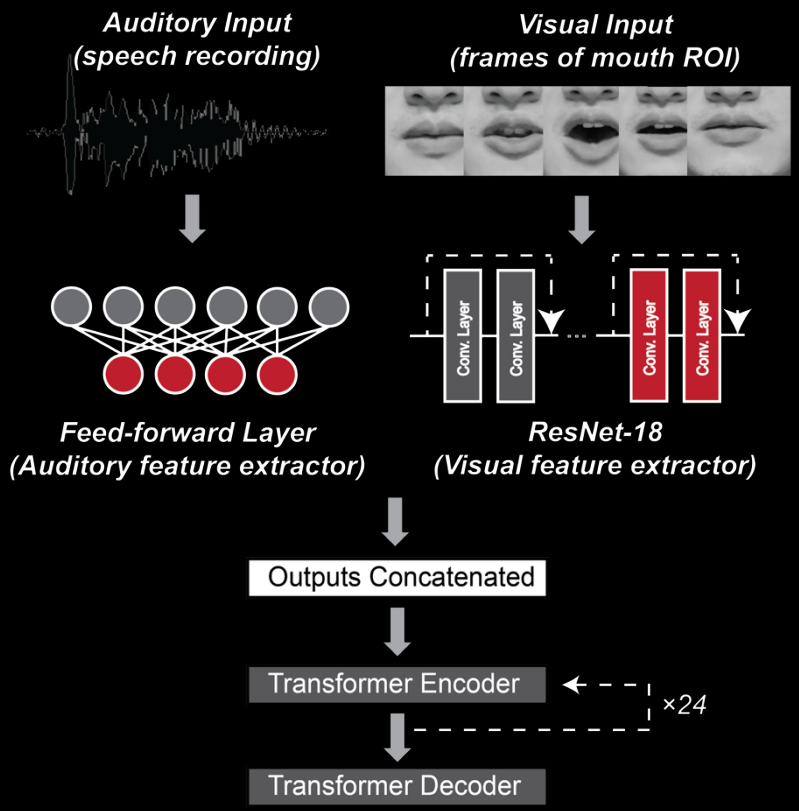
Predict perception of arbitrary combinations of A and V speech



Do DNNs “perceive” the McGurk effect?



DNN Model: AVHuBERT



- Audiovisual Hidden-unit Bidirectional Encoder Representations from Transformers
- Meta Corp.: Shi et al., Learning audio-visual speech representation by masked multimodal cluster prediction (ICLR 2022)

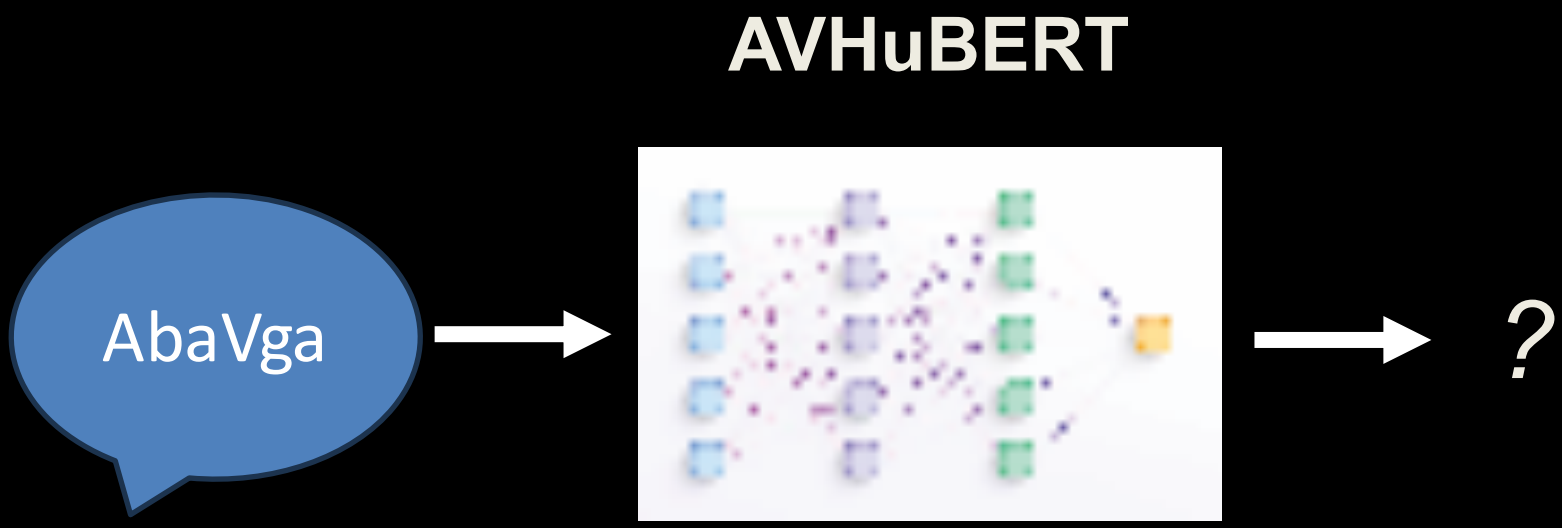
Model features

- Off-the-shelf model downloaded from Github

  no additional training

- trained on hundreds of hours of *congruent* speech. How will it “perceive” *incongruent* speech such as McGurk syllables?

Does AVHuBERT “perceive” the McGurk effect?

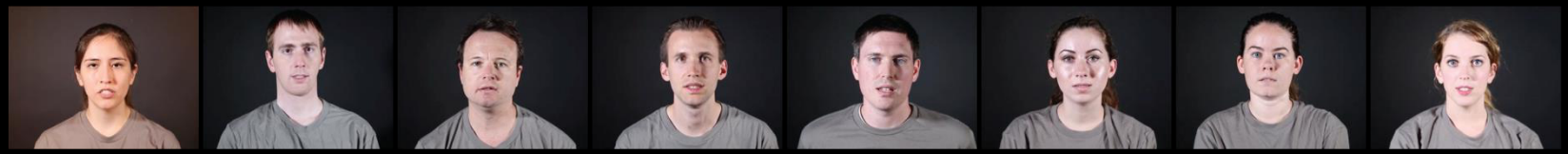


Sample AbaVga stimulus

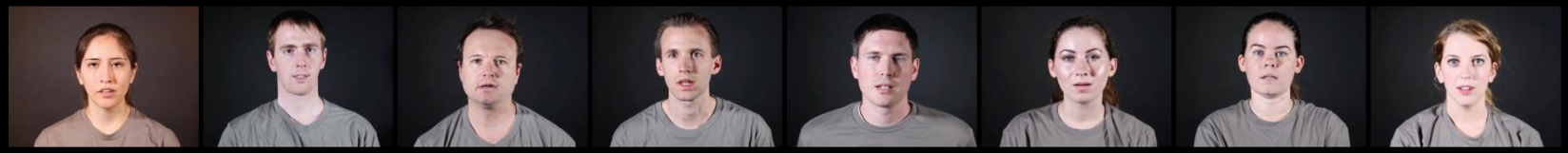
8-talker stimulus set

Congruent: AbaVba

Congruent: AdaVda



Congruent: AgaVga



McGurk: AbaVga

+ visual-only, auditory-only: ba, da, ga

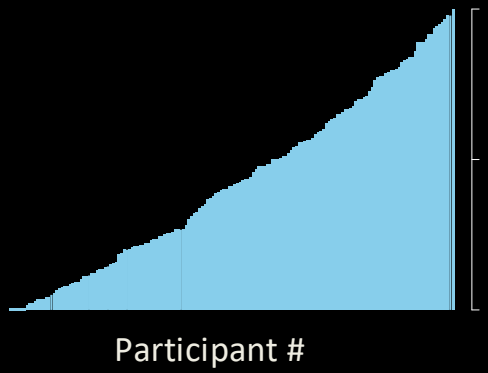
Methods

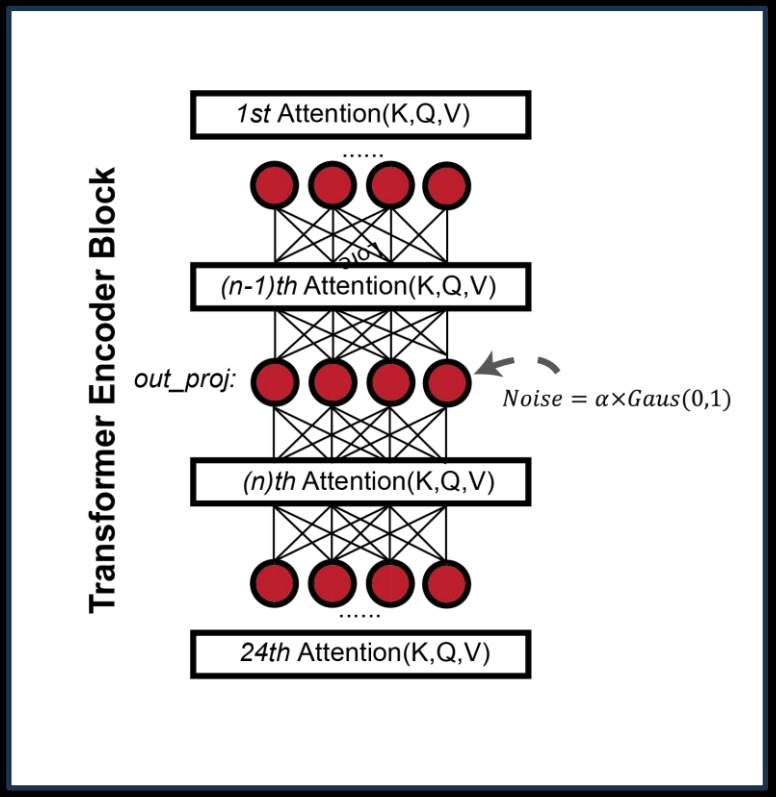
- For each input token, AVHuBERT ranks 5 most likely words (always provides a word response because trained on words)
- Classify first phoneme of each word as fusion; auditory; visual; other

for comparison with human studies with the same set of response choices

- Calculate likelihood and convert to percentage

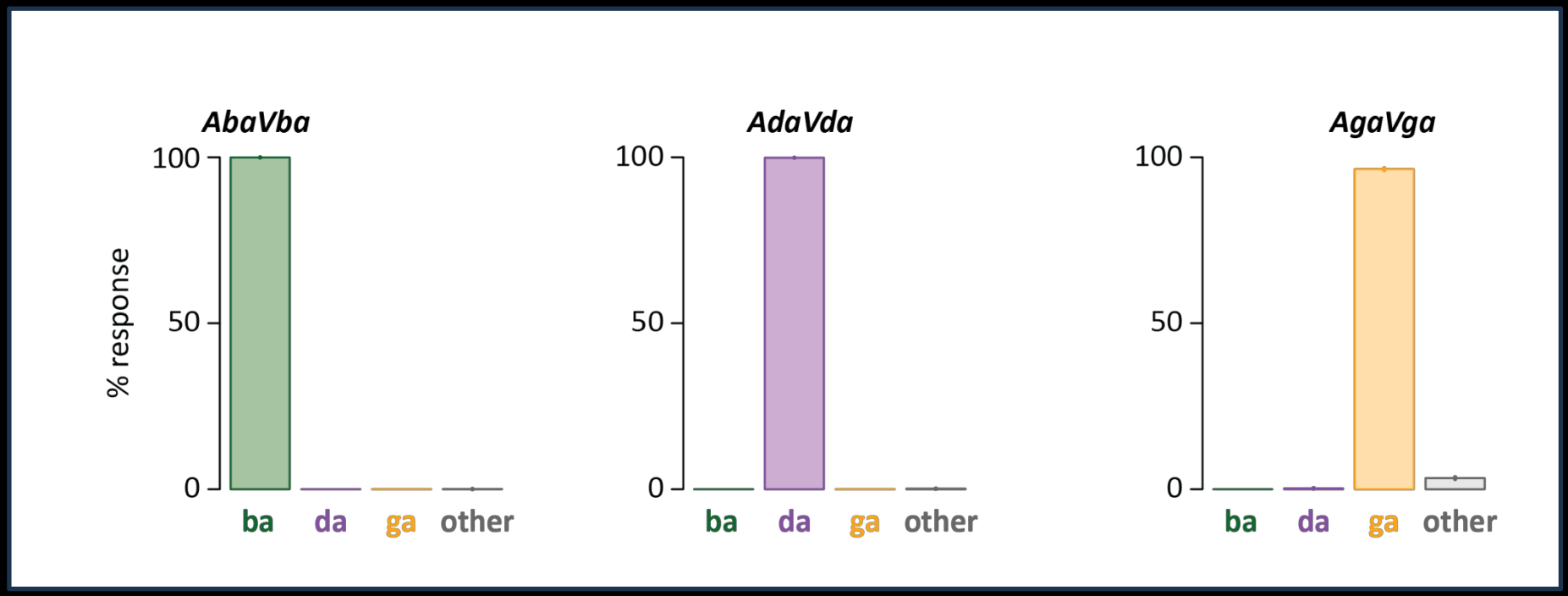
Methods II



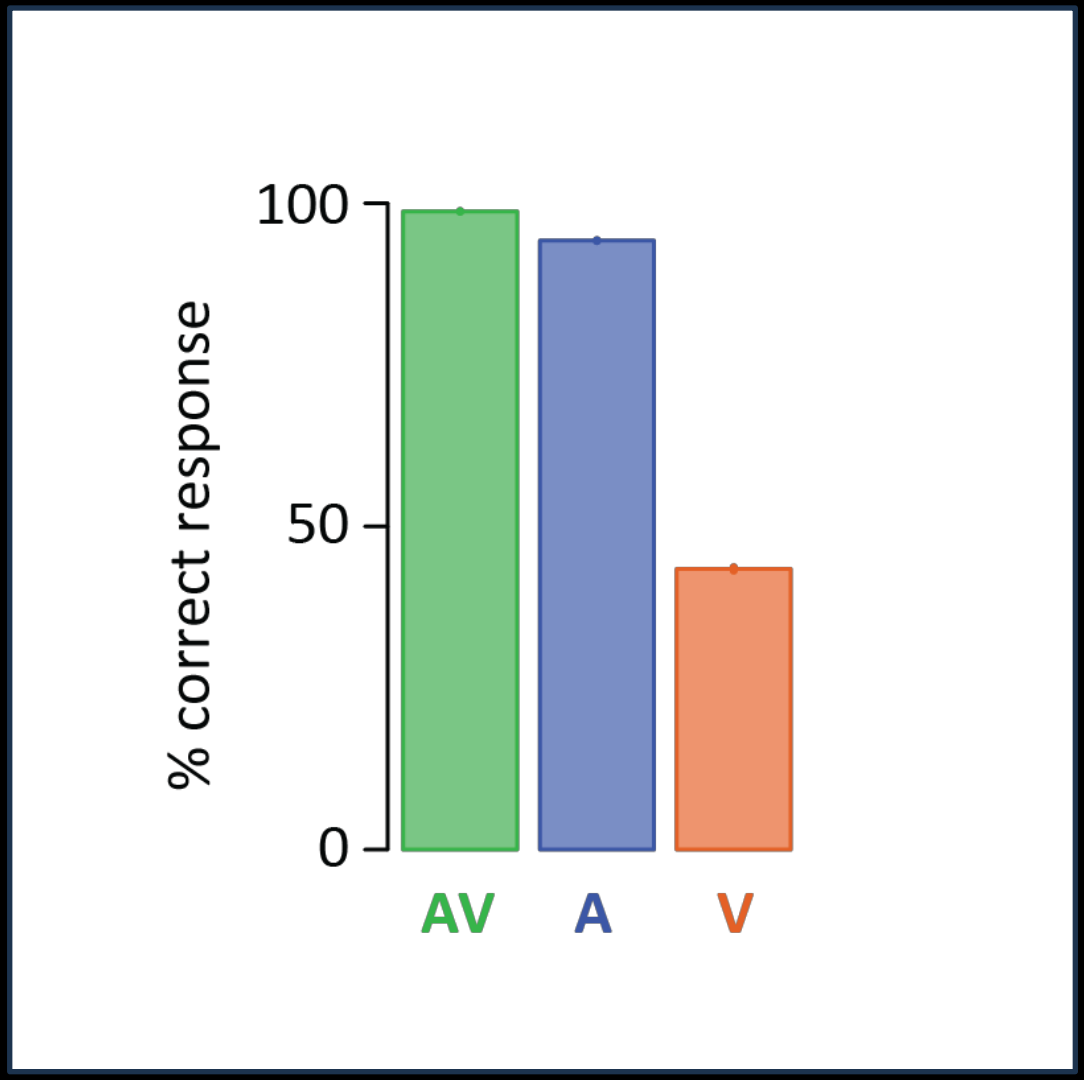


- Create 100 AVHuBERT variants to model human variability
- 100 random Gaussian arrays, added to learned weights of a layer important for audiovisual integration

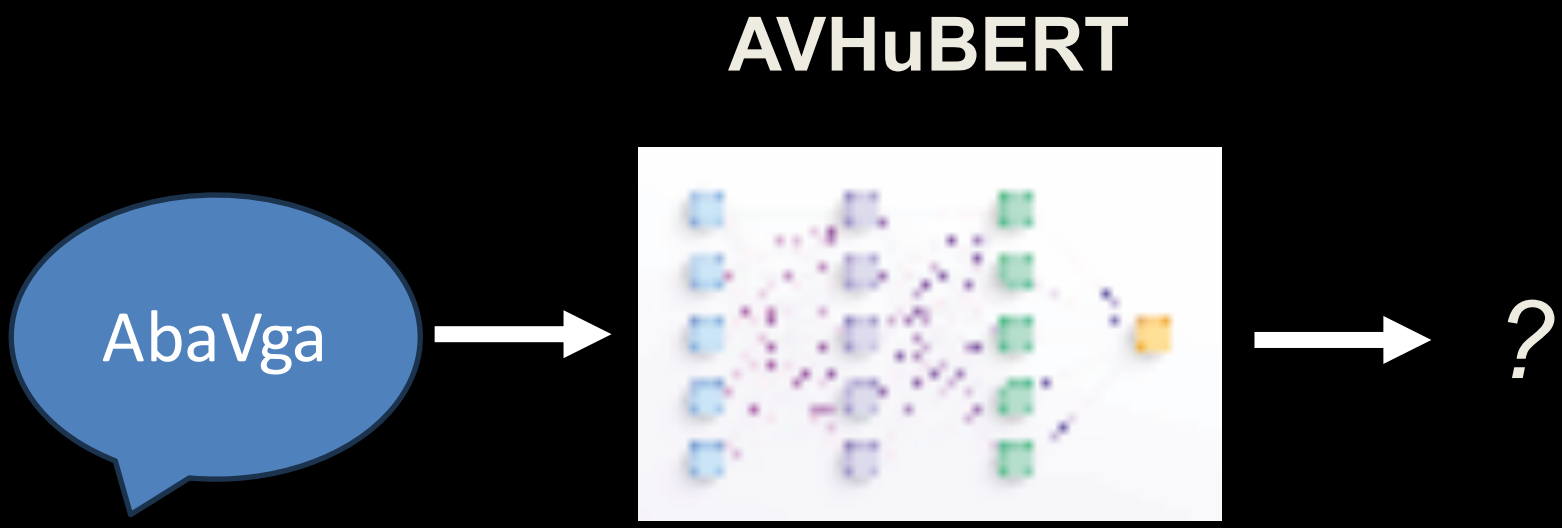
Results: congruent AV syllables, 8 talkers



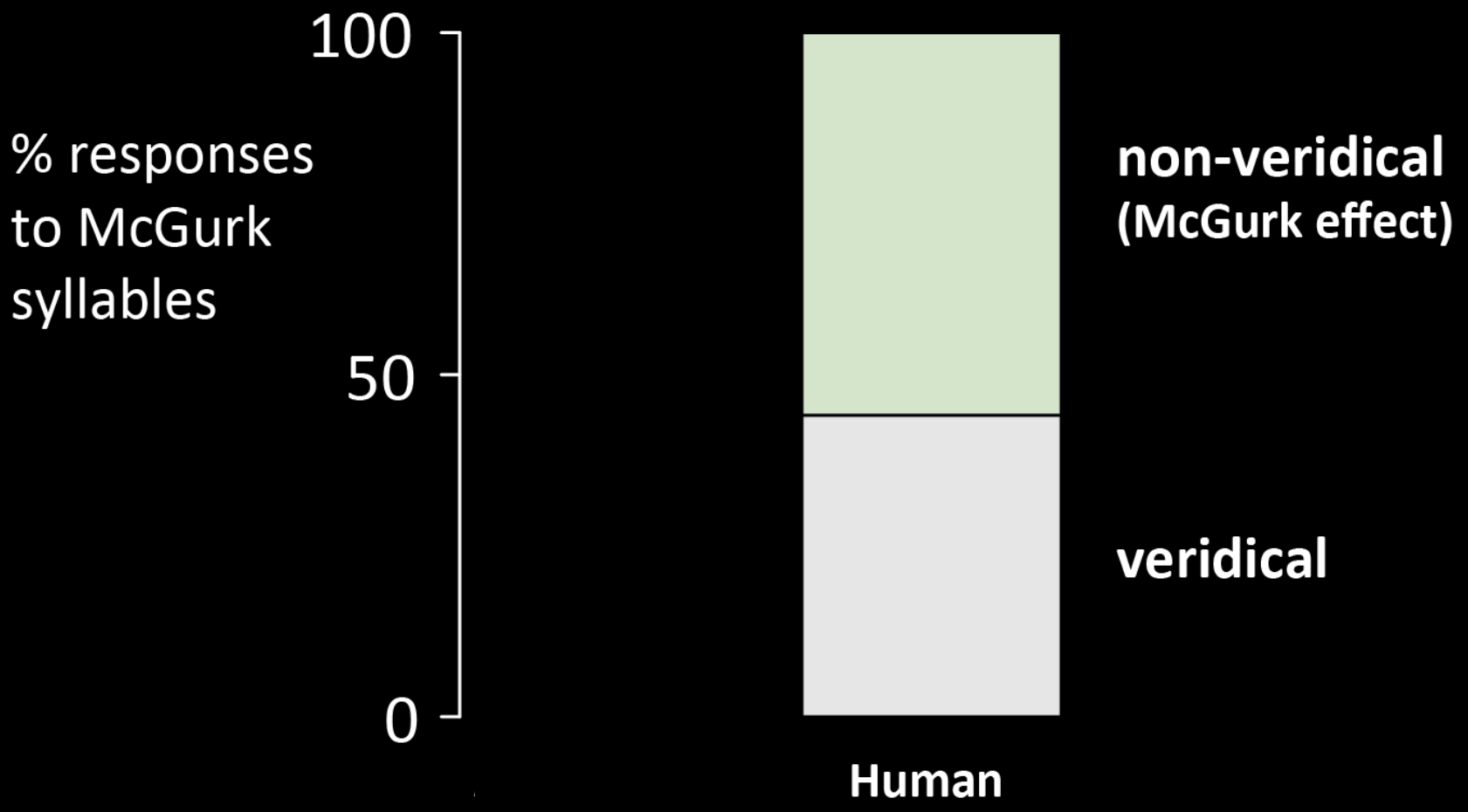
Results: A, V, congruent AV syllables



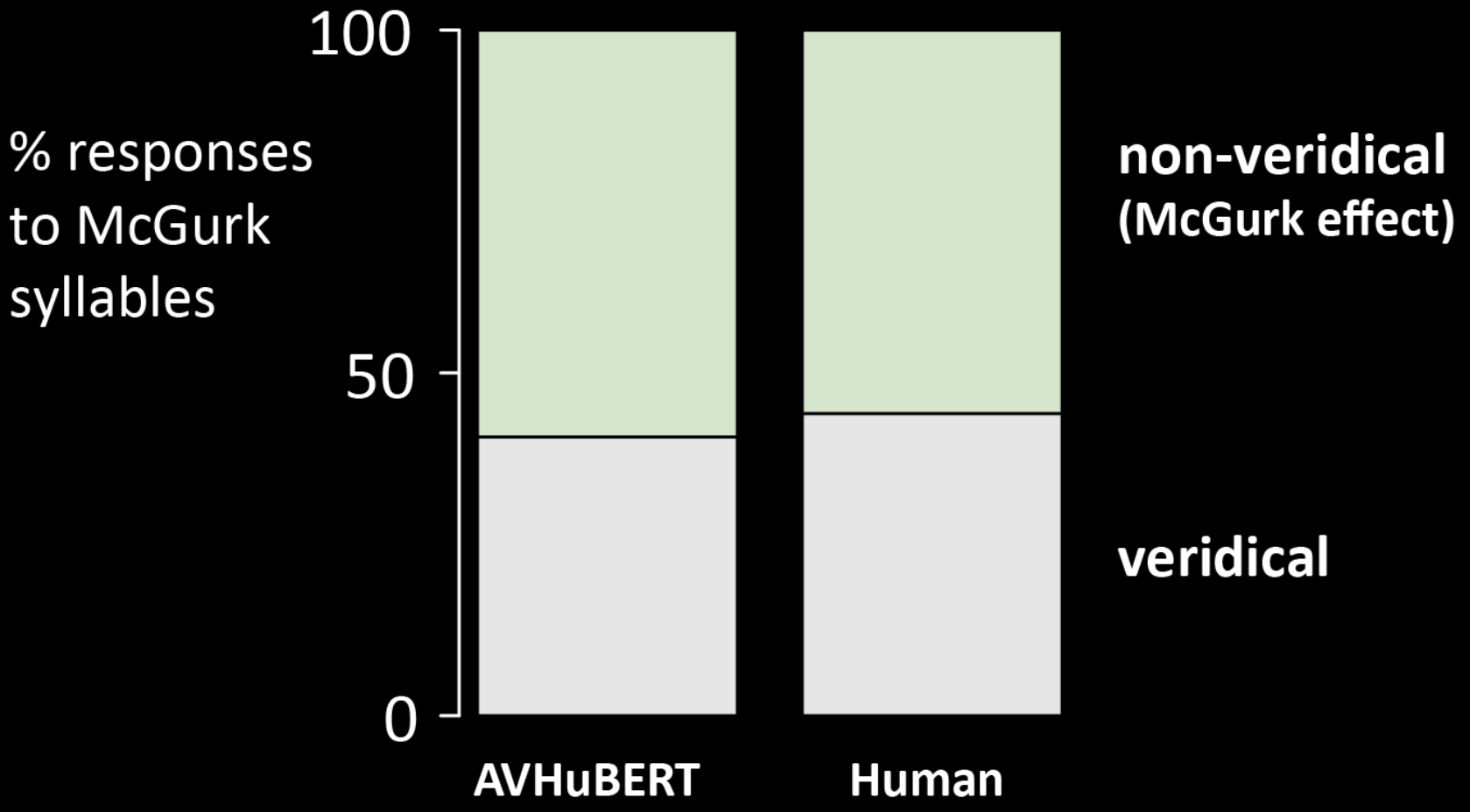
Does AVHuBERT “perceive” the McGurk effect?



Results: McGurk syllables

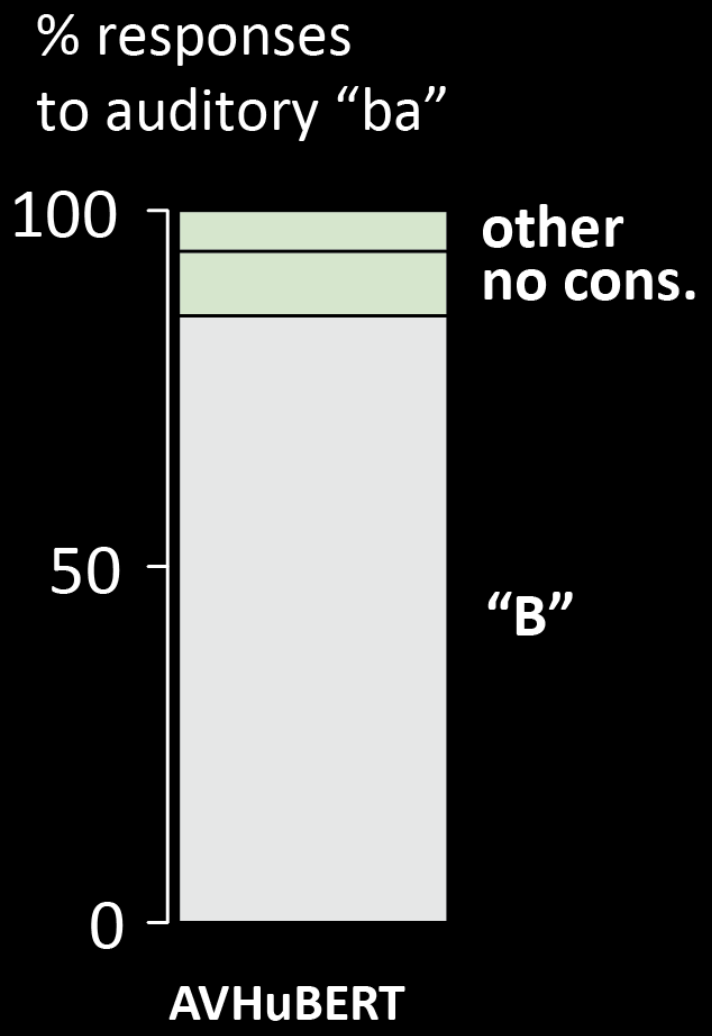


Results: McGurk syllables (AbaVga)

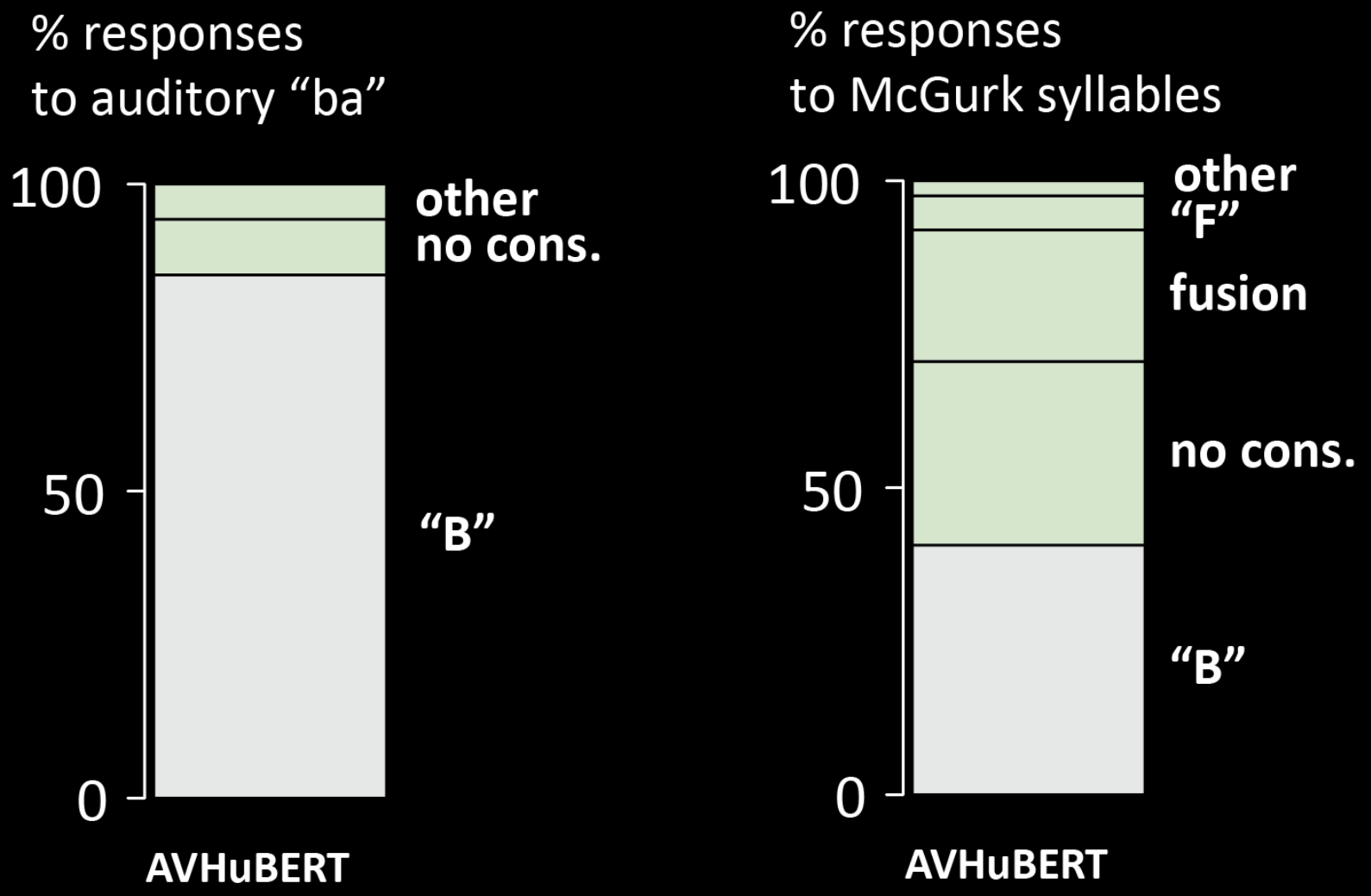


Does AVHuBERT “perceive” the McGurk effect? YES

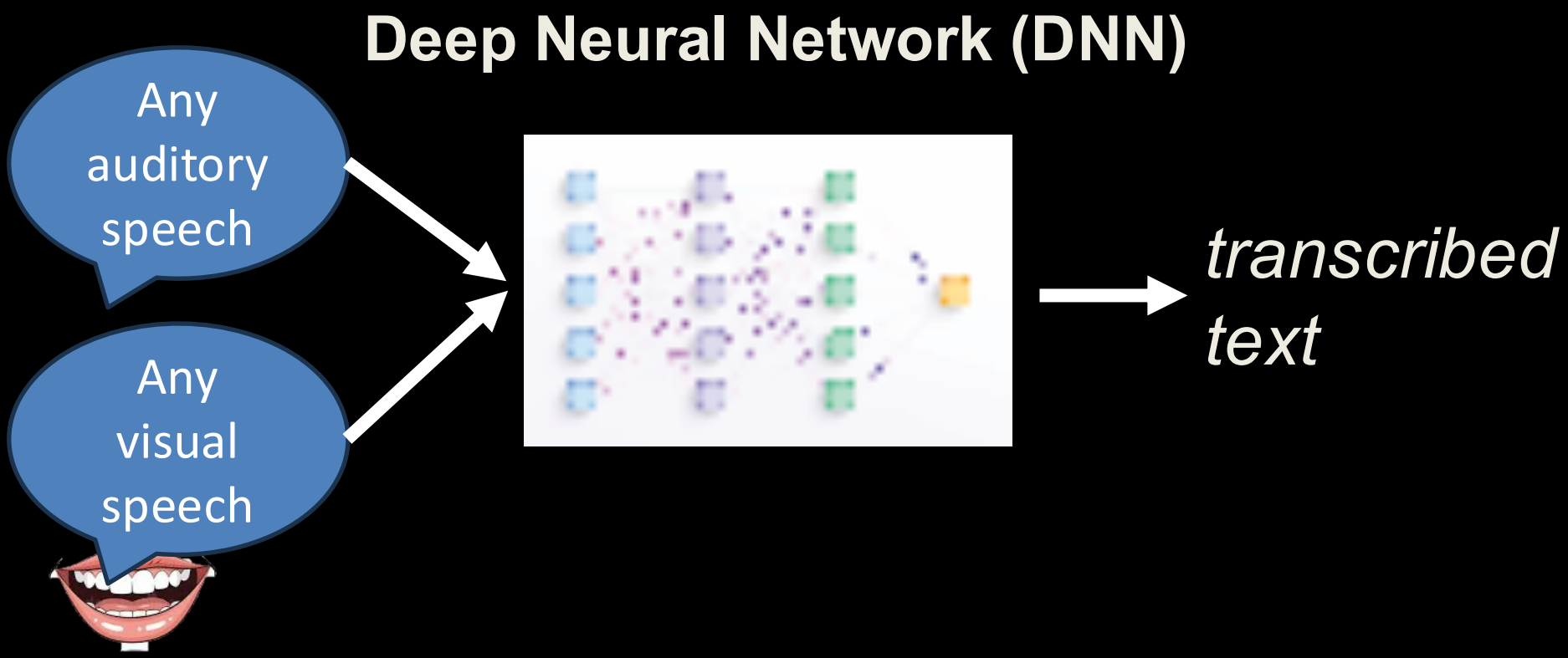
Response to auditory "ba" (Aba)



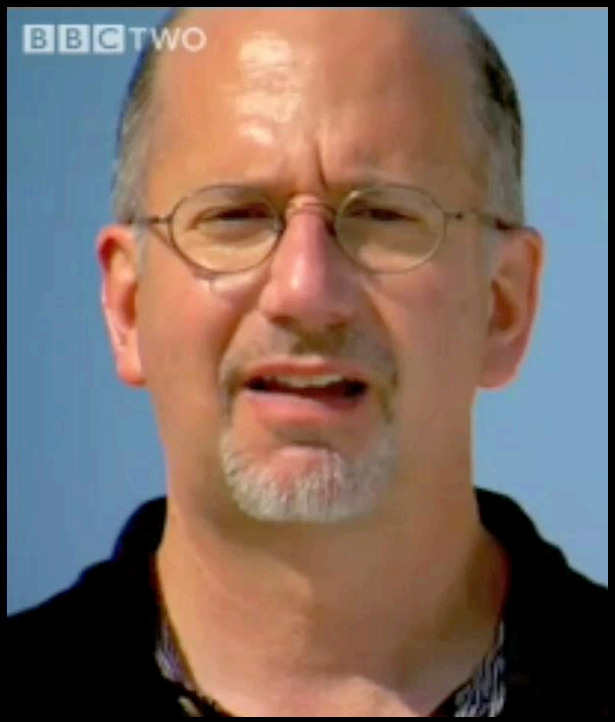
Response to auditory "ba" *vs.* response to AbaVga



Predict perception of arbitrary combinations of A and V speech



Auditory “ba” + Visual “fa”



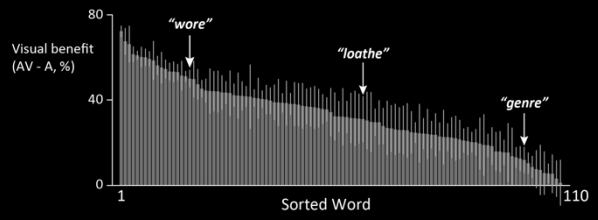
AVHuBERT:
>95% “fa” reports

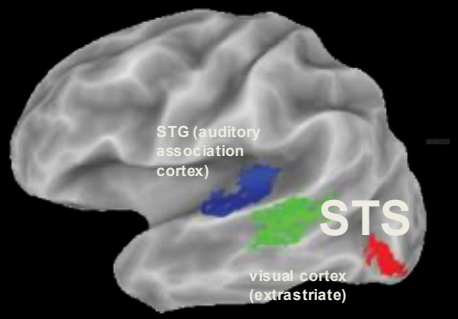
Section summary

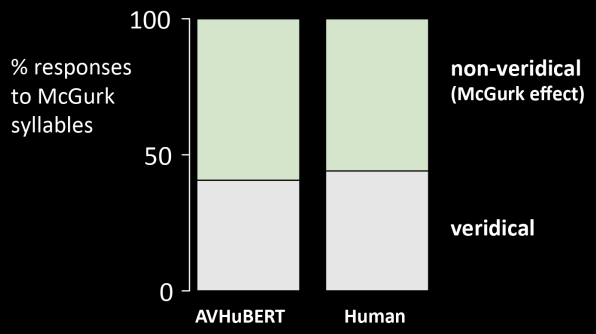
- DNNs trained for audiovisual integration display many similarities (and some differences) from human observers
- DNNs may provide a useful tool for exploring the neural substrates of audiovisual speech perception

Questions?

Overall summary







1) Large differences across words/phonemes in visual benefit
2) A network of brain areas, especially STS, for audiovisual speech
3) DNNs reproduce some aspects of human audiovisual speech perception

People who did the work

former trainees who contributed to the work presented today


Yue Zhang


Debshila Basu Mallick


Patrick Karas


Johannes Rennig


Katie Lee


Brenna Argall


Kira Wegner-Clemens


Brian Metzger

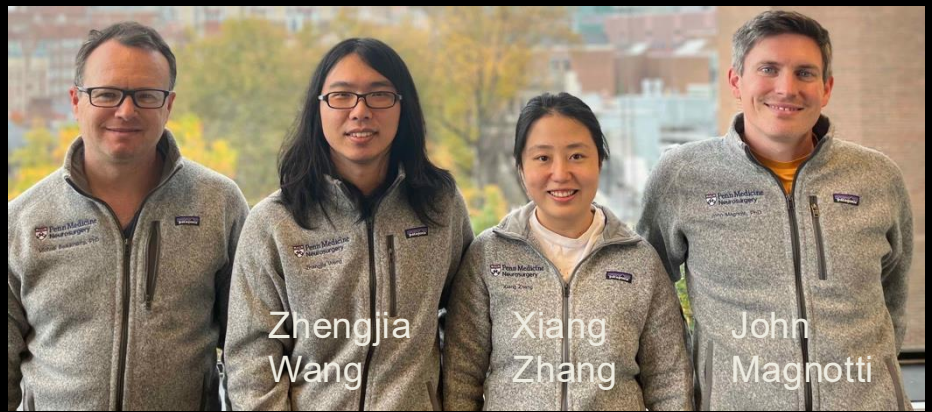

Nafi Yasar


Siavash Pasalar


Anastasia Lado

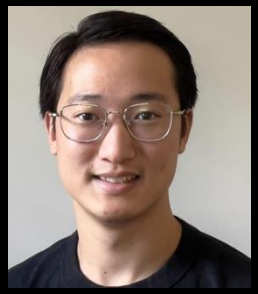

IzzieYu

current lab members


Zhengjia Wang

Xiang Zhang

John Magnotti


Haotian Ma


Ilaina Edelstein

Acknowledgements

michael.beauchamp@pennmedicine.upenn.edu

reprints, stimuli and code at

beauchamplab.com

NIH National Eye Institute (NEI)
Research Today...Vision Tomorrow

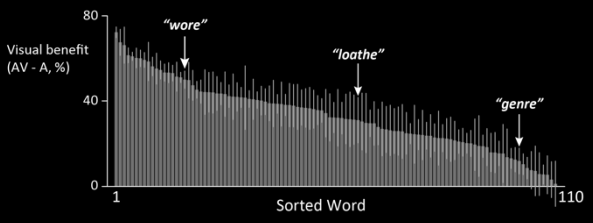
UNITED STATES
DEPARTMENT OF VETERANS AFFAIRS

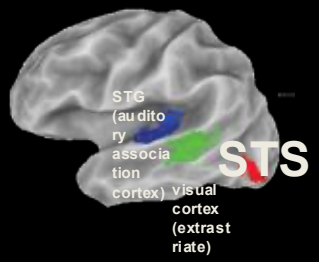
BRAIN Initiative®

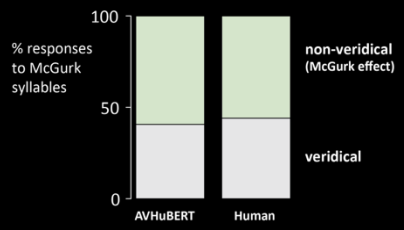
DFG Deutsche Forschungsgemeinschaft

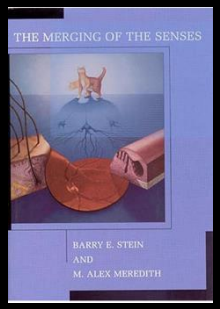
National Institute of
Neurological Disorders and Stroke
National Institutes of Health

Discussion









- Behavior
  - *phoneme-overlap analysis*
- Neural substrates
  - *audiovisual integration in STS*
- Deep neural networks
  - *behavior and neural models*
- Multisensory integration
  - understudied but important!