Dissecting neural computations of the human auditory pathway - from hypothesis-driven encoding to deep neural nets

Yuanning Li

Neuroimaging Methods Workshop

City University of Hong Kong

December 11, 2022

yuanningli@gmail.com
https://yuanningli.github.io/

上海科技大学 ShanghaiTech University

Demo Code

- GitHub: https://github.com/yuanningli/neural_encoding_demo
- QR code:

yuanningli@gmail.com
https://yuanningli.github.io/

Acknowledgments

Edward Chang (UCSF)

Claire Tang (UCSF)

Gopala Anumanchipalli (Berkeley)

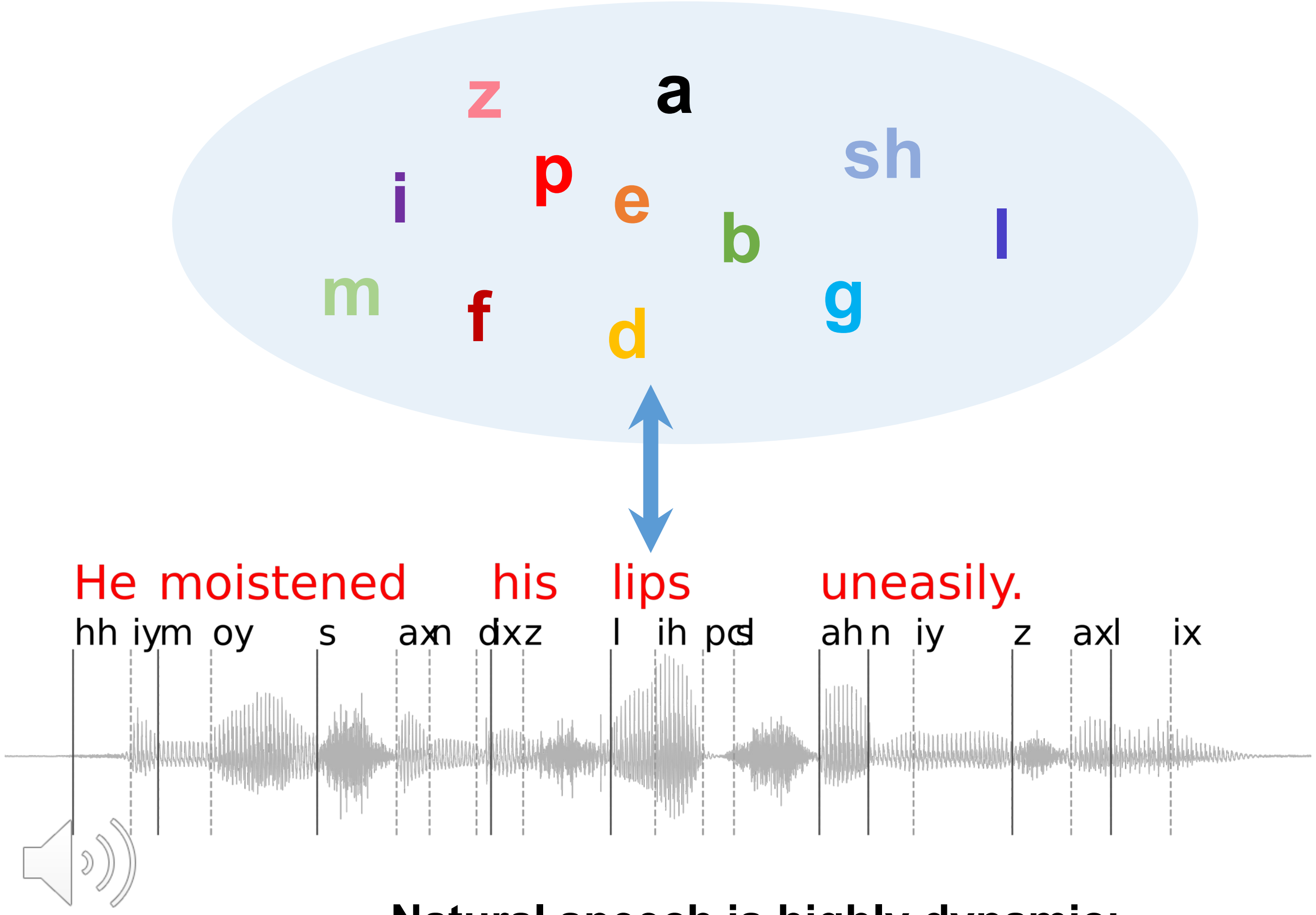
Abdelrahman Mohamed (Meta AI)

Laurel Carney (Rochester)

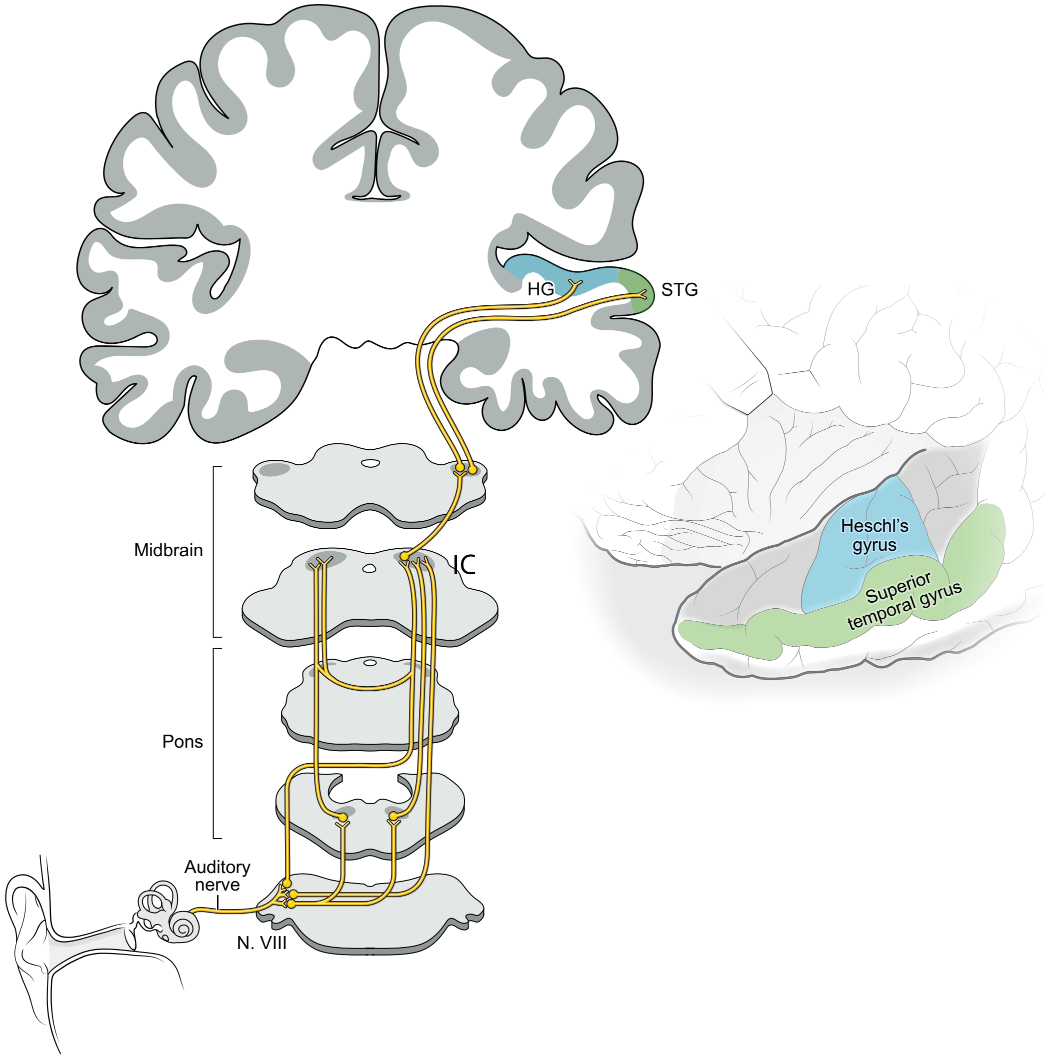
Jinsong Wu (Huashan)

Junfeng Lu (Huashan)

Transformation of speech sound into phonetic units



Natural speech is highly dynamic:
~150 words per minute,
~15-20 phonemes per second



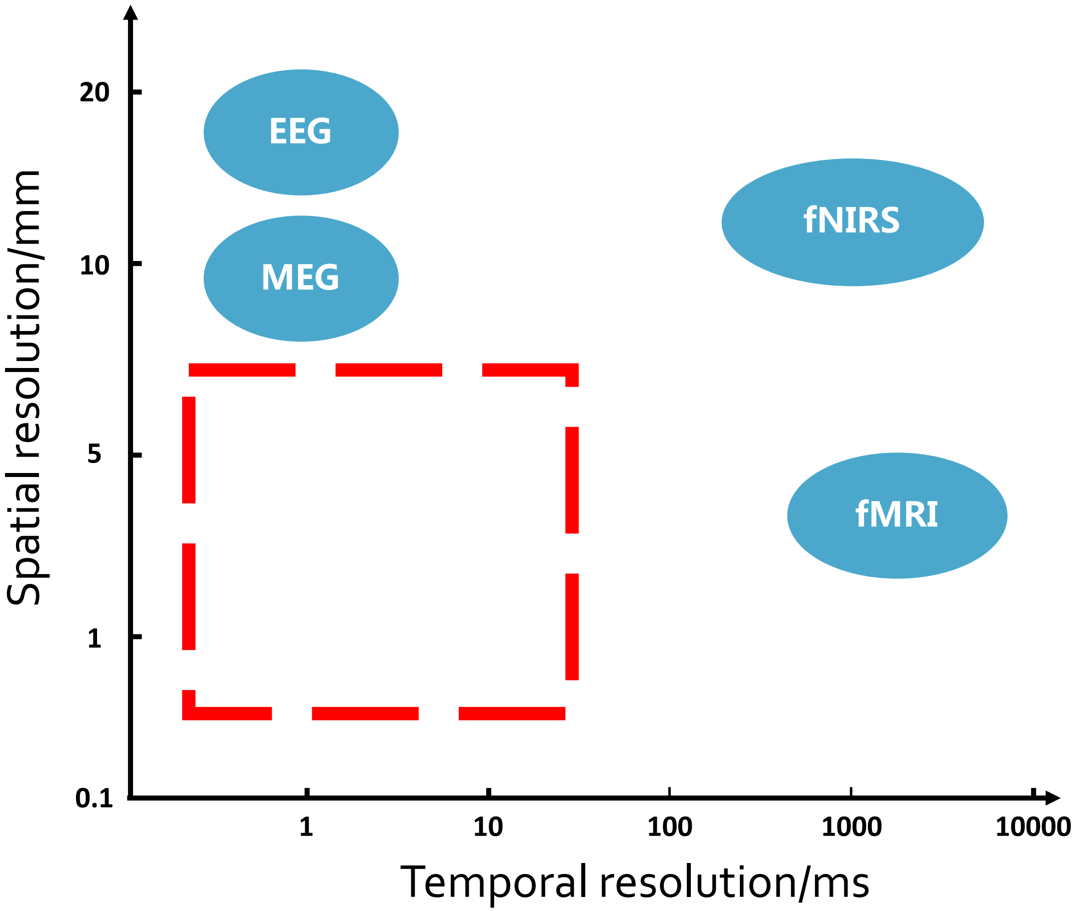
Spatiotemporal resolution of imaging modalities

EEG

fNIRS

fMRI

MEG

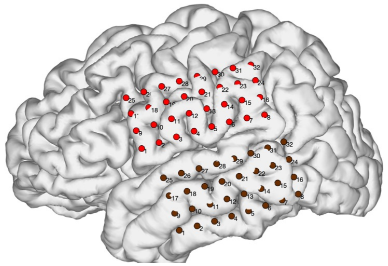


- Cannot achieve high spatiotemporal resolution
- Noninvasive methods have low SNR

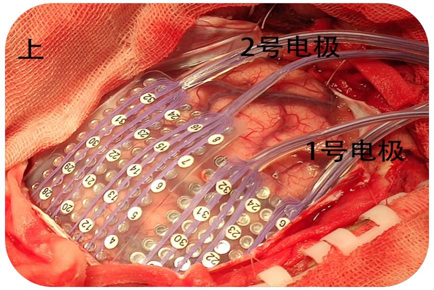
Spatiotemporal resolution of Electrocorticography (ECoG)

Intracranial EEG (ECoG)

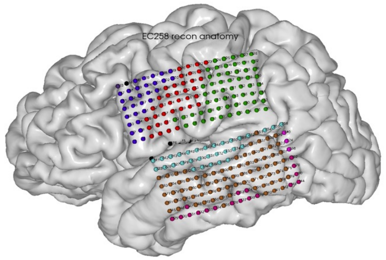
Awake surgery



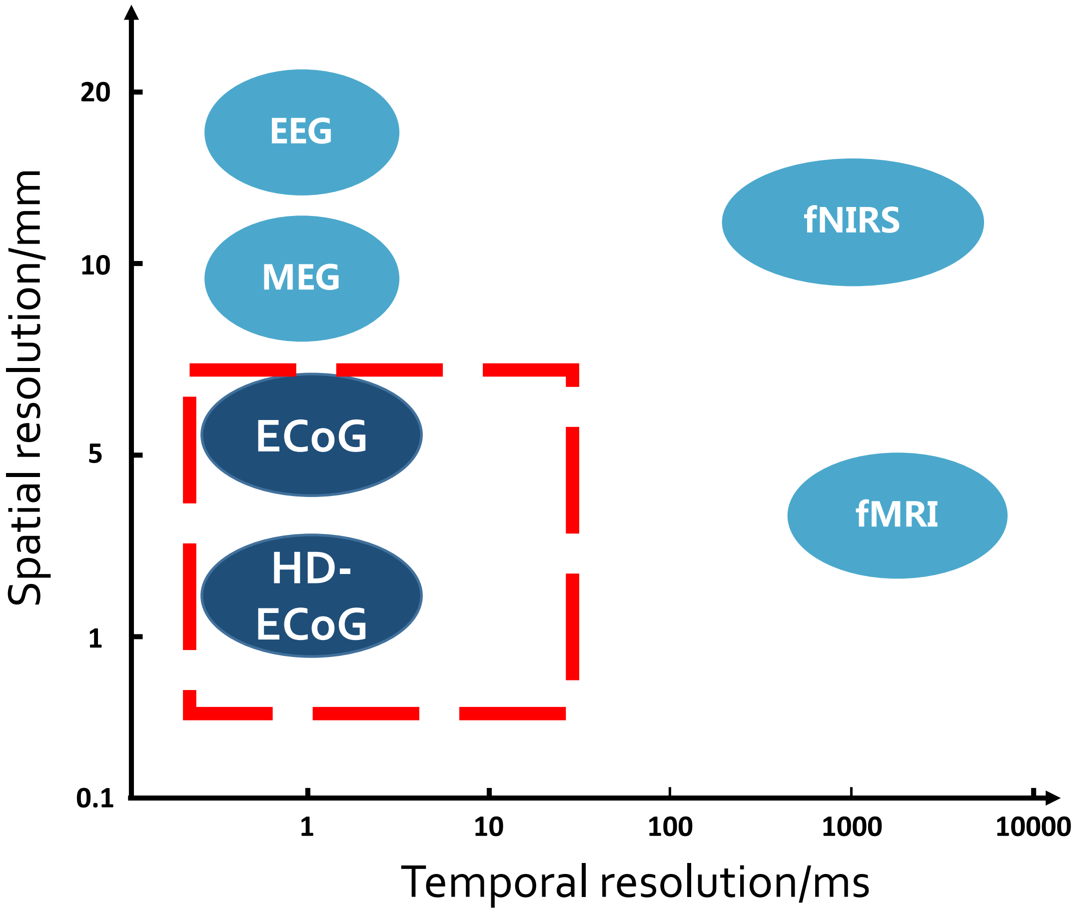
ECoG



High density ECoG



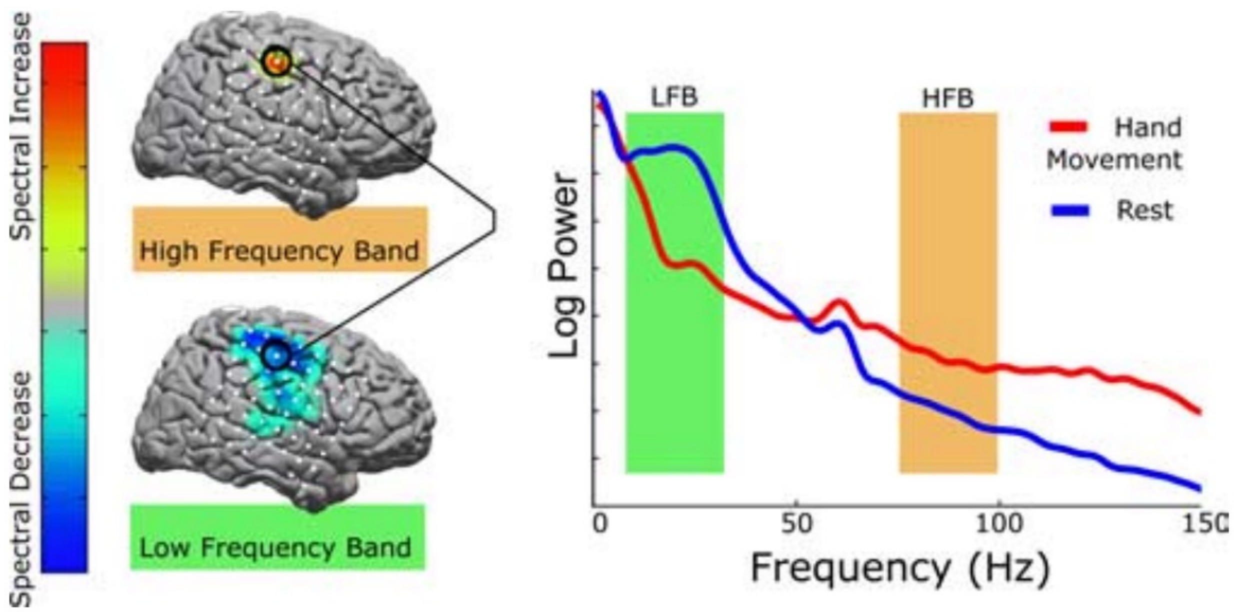
High density ECoG



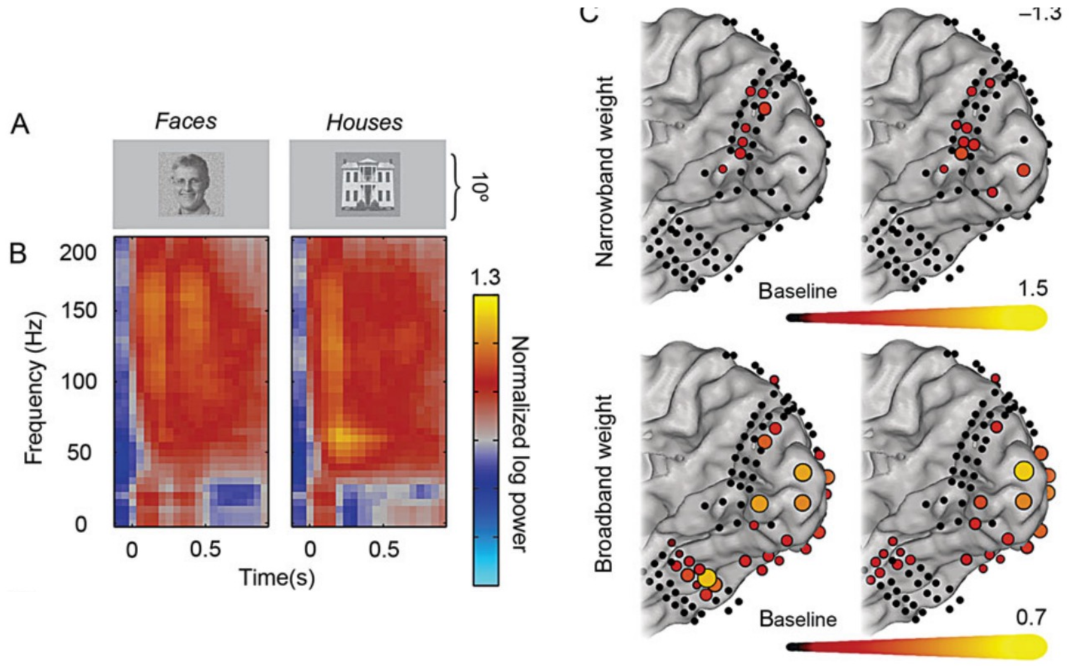
- **High spatial (~1mm , 256 channels in 5.5*5.5cm^2) and temporal (ms) resolution**
- **One of the highest SNR methods for human *in vivo* neural recording**

Neural electrophysiology signals recorded by ECoG

- Broadband high-gamma response in sensory and motor cortex
  - ~70-150Hz broadband signal
  - Reflecting local neuronal activity



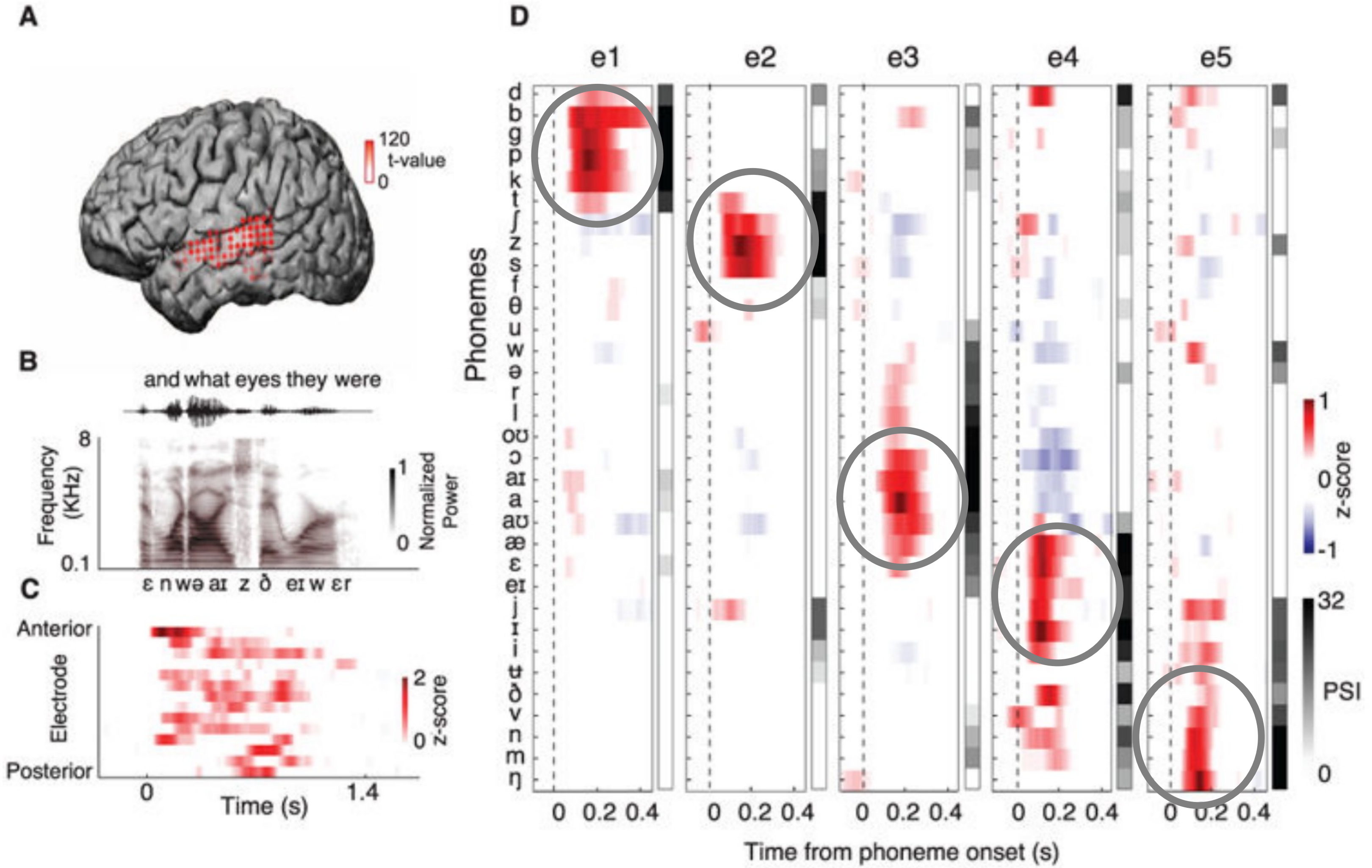
Miller *et al. J. Neurosci.* 2007



Hermes *et al. Cereb. Cortex* 2015

Superior temporal gyrus (STG) codes for phonetic features.

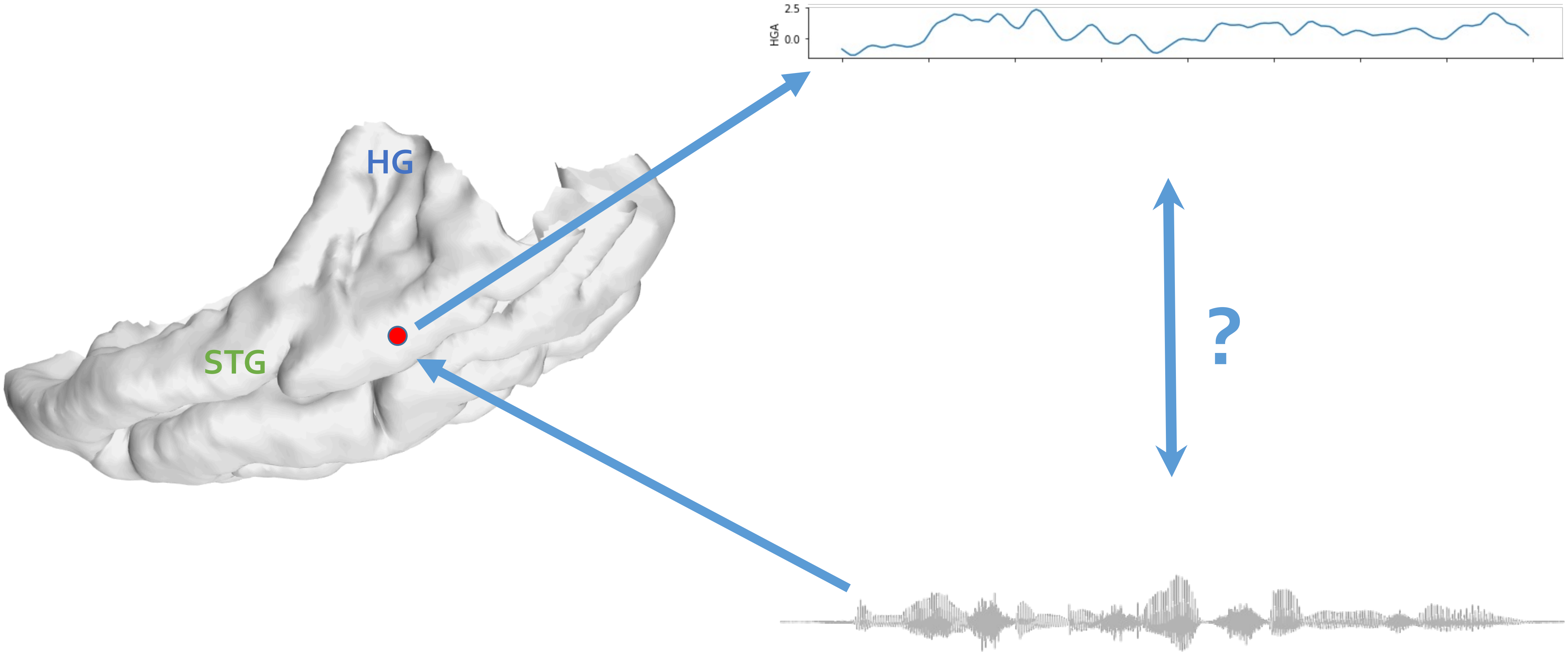
- Different electrodes tune to different phonetic features --- A spatial code for acoustic-phonetic features



Mesgarani et al., *Science*, 2014

A neural encoding problem

- What are the features in speech that drive neural activity in cortex?



Marr’s three levels of analysis

David Marr’s three levels

- **Computational level**
  - cognitive task and problem
- **Algorithm level**
  - information representation and transformation
- **Implementation level**
  - implementation through interactions between basic elements

Neural coding

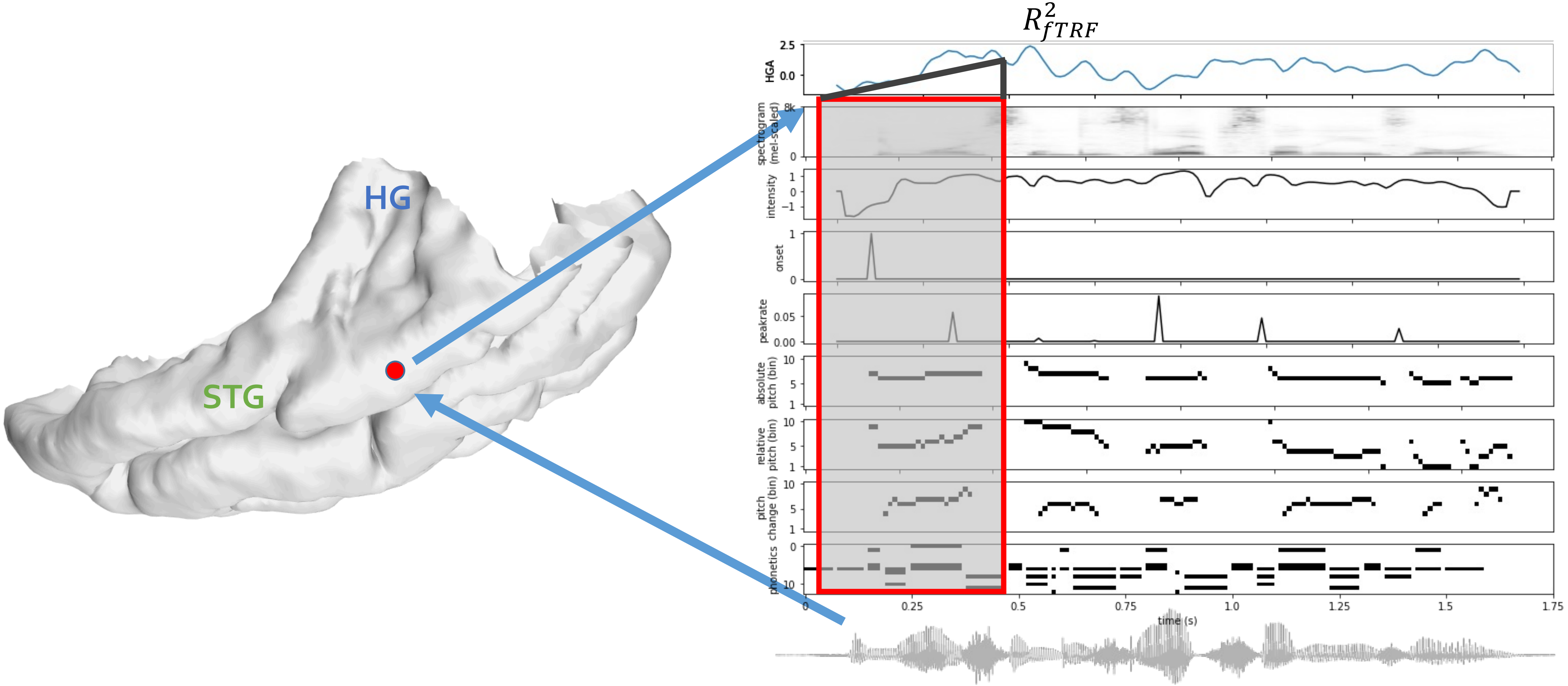
- Fit the underlying neural activity
- Verify theoretical hypotheses

Computational model

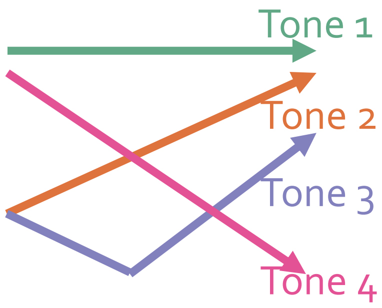
Marr 1982

Hypothesis-driven linear encoding models

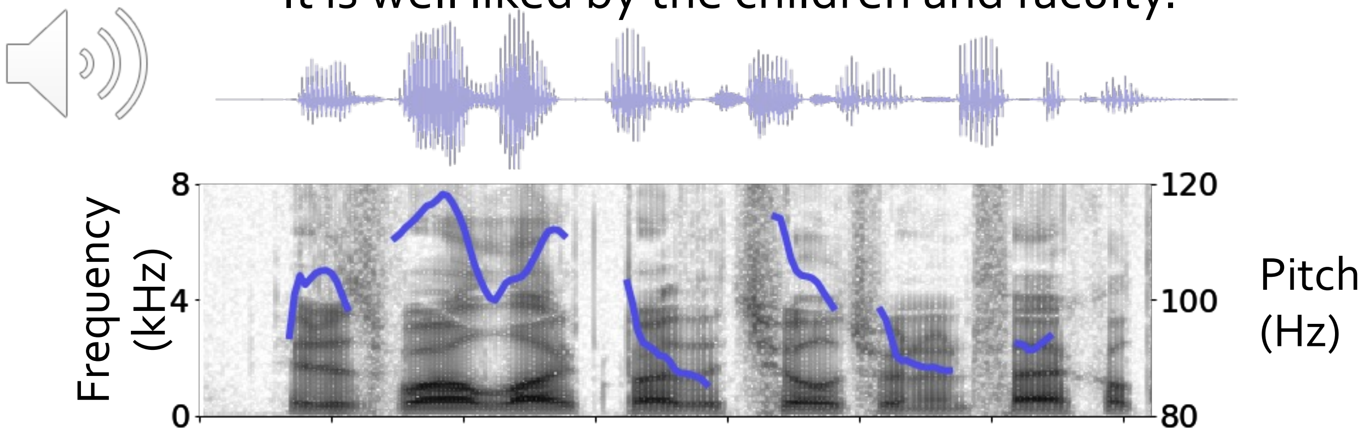
- Linear temporal receptive field model reveals neural coding for distinct speech features in the human auditory cortex



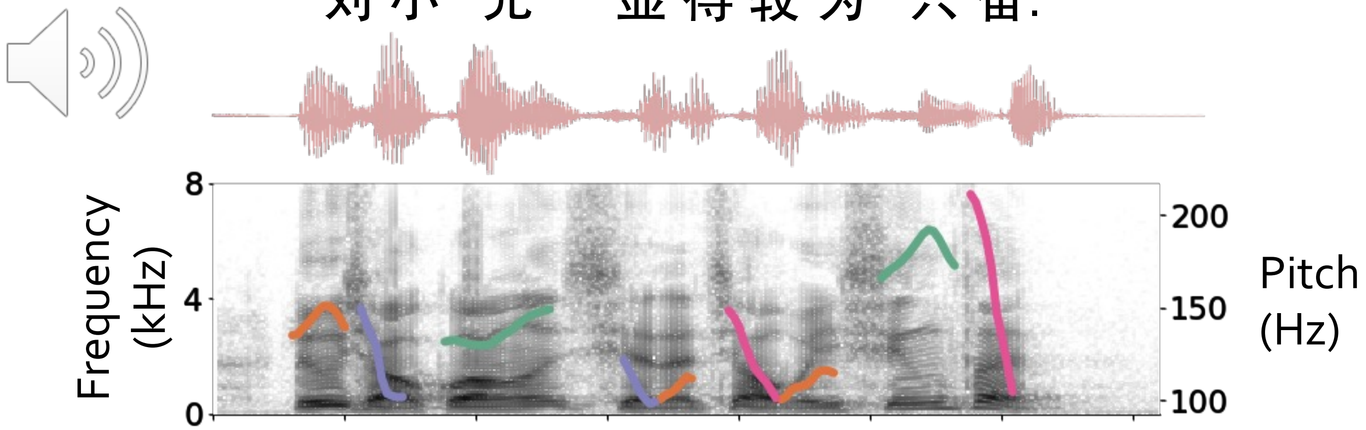
Tonal languages use pitch to distinguish word meanings

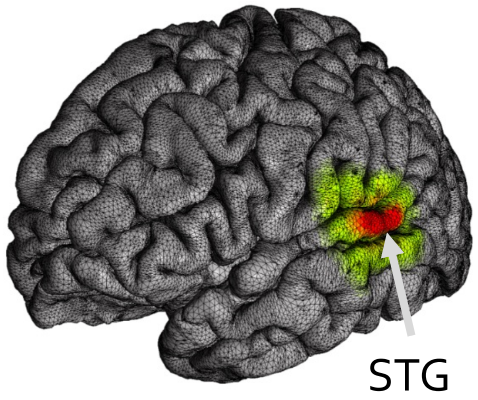


"It is well liked by the children and faculty."



"刘 小 光 显 得 较 为 兴 奋."

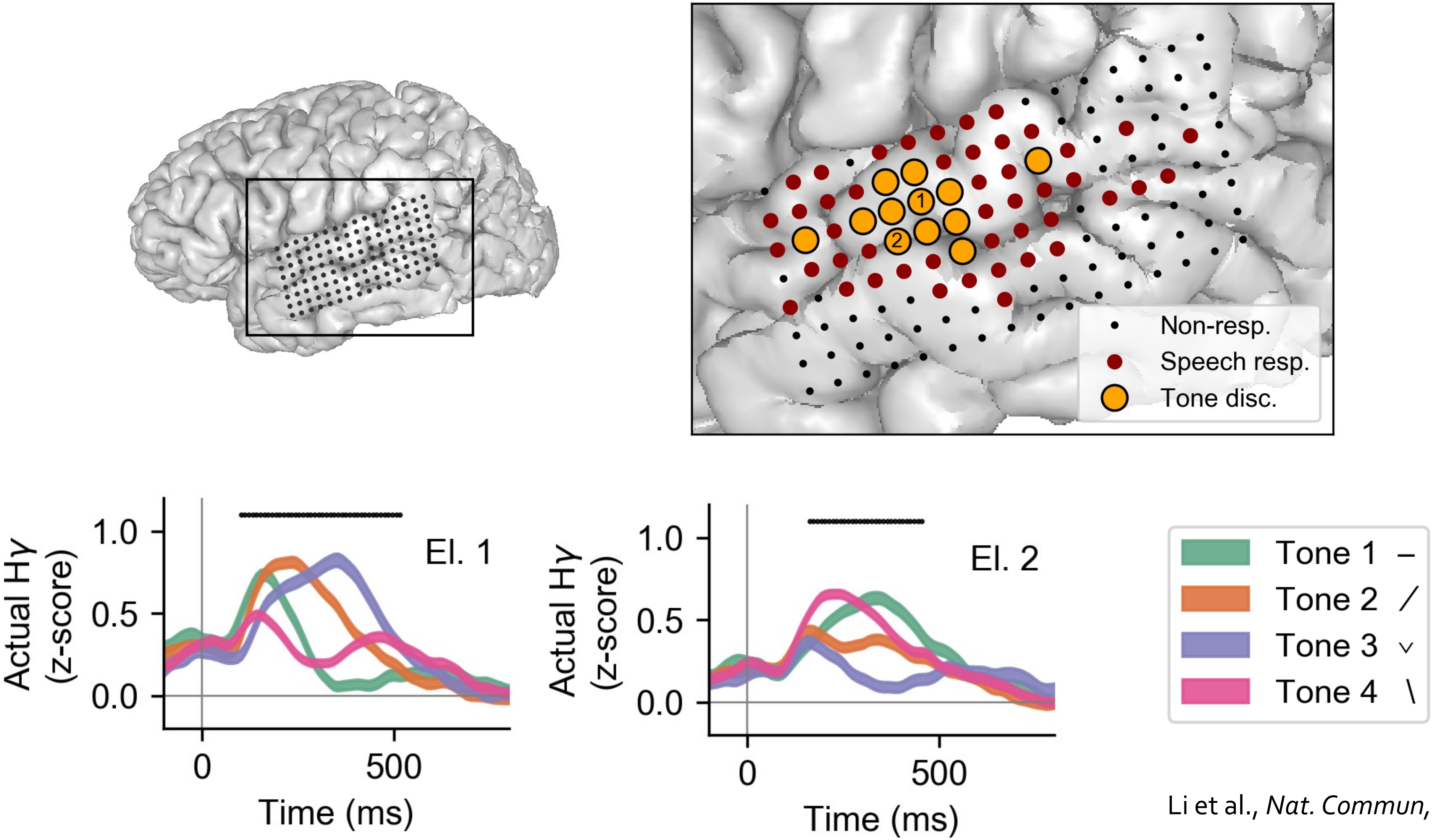




Research questions

- What features are encoded in STG in service of lexical tone representation?
  - Lower-level acoustic cues?
  - Complex intermediate features?
  - Abstract tone category?
- Is the neural computation underlying lexical tone perception language-specific?
  - Are the encoding properties shared across languages and across listeners with different language experiences?

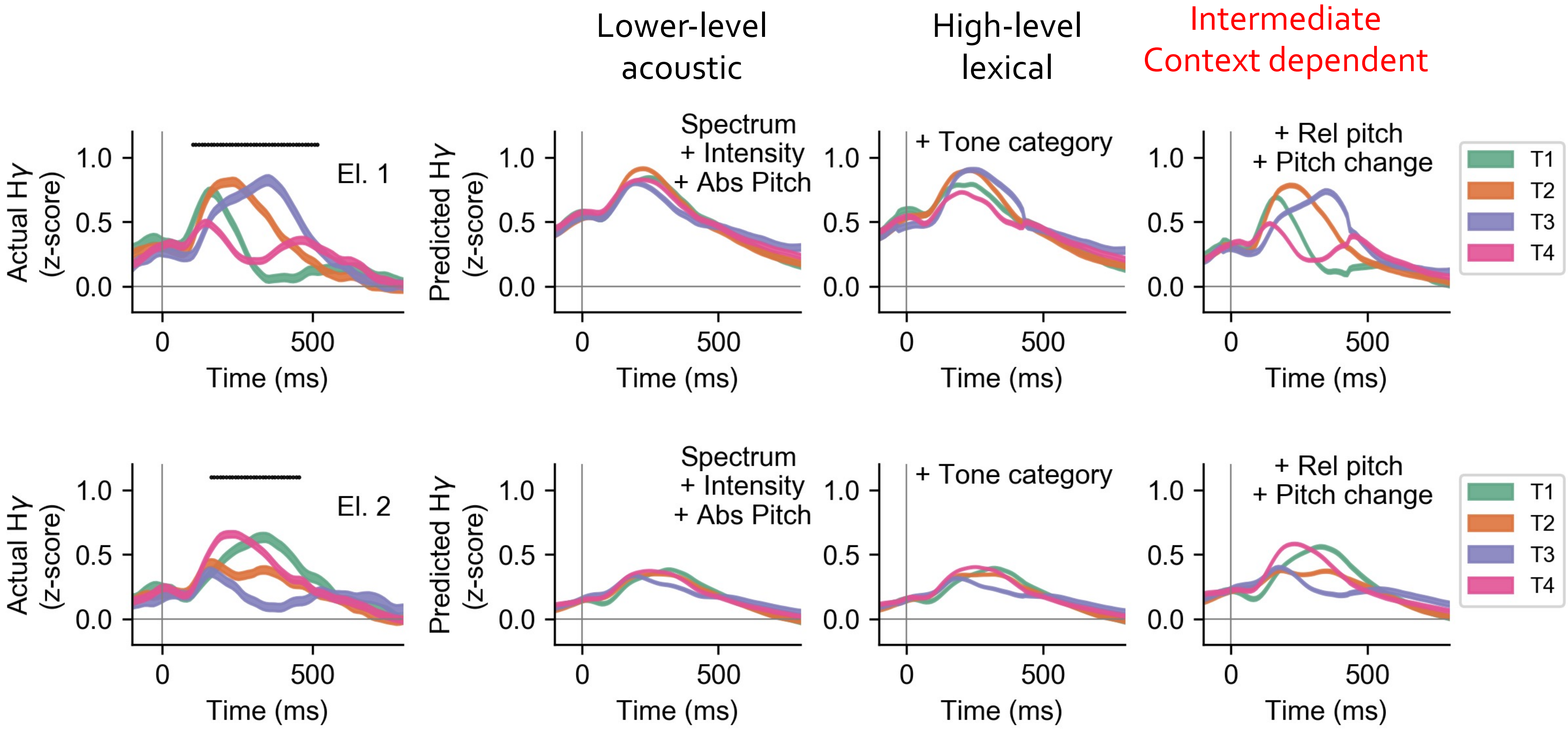
Lexical tones in continuous Mandarin speech evoke differential neural responses in discrete populations in STG



Li et al., *Nat. Commun*, 2021

The differential neural responses are mainly driven by speaker-normalized pitch features, rather than discrete tone category

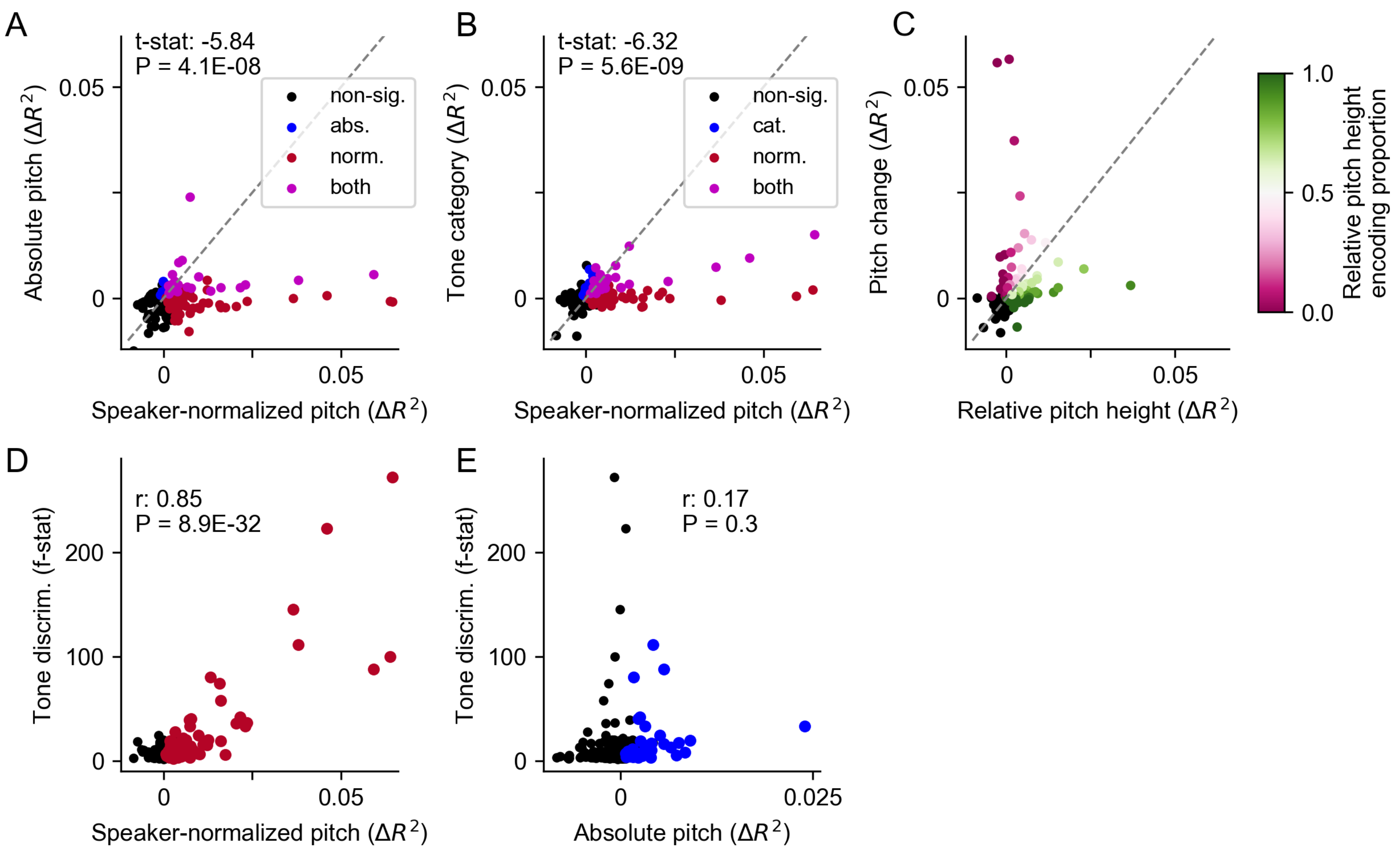
上海科技大学 ShanghaiTech University



Li et al., *Nat. Commun.*, 2021

The differential neural responses are mainly driven by speaker-normalized pitch features, rather than discrete tone category

上海科技大学 ShanghaiTech University



Li et al., *Nat. Commun.*, 2021

Research questions

上海科技大学 ShanghaiTech University

- What features are encoded in service of lexical tone representation?
  - Lower-level acoustic cues?
  - Complex intermediate features?
  - Abstract tone category?

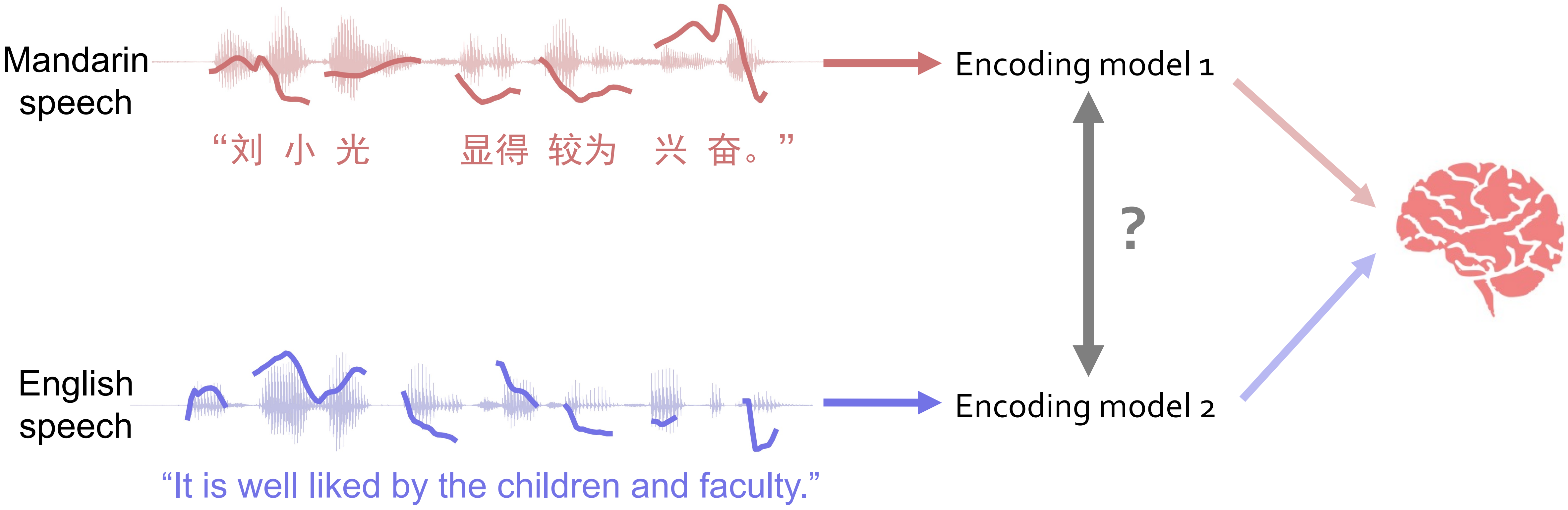
Research questions

- What features are encoded in service of lexical tone representation?
  - Lower-level acoustic cues?
  - Complex intermediate features: speaker-normalized pitch (height and change)
  - Abstract tone category?

- Is the neural computation underlying lexical tone perception language-specific?
  - Is the encoding properties shared across languages and across listeners with different language experiences?

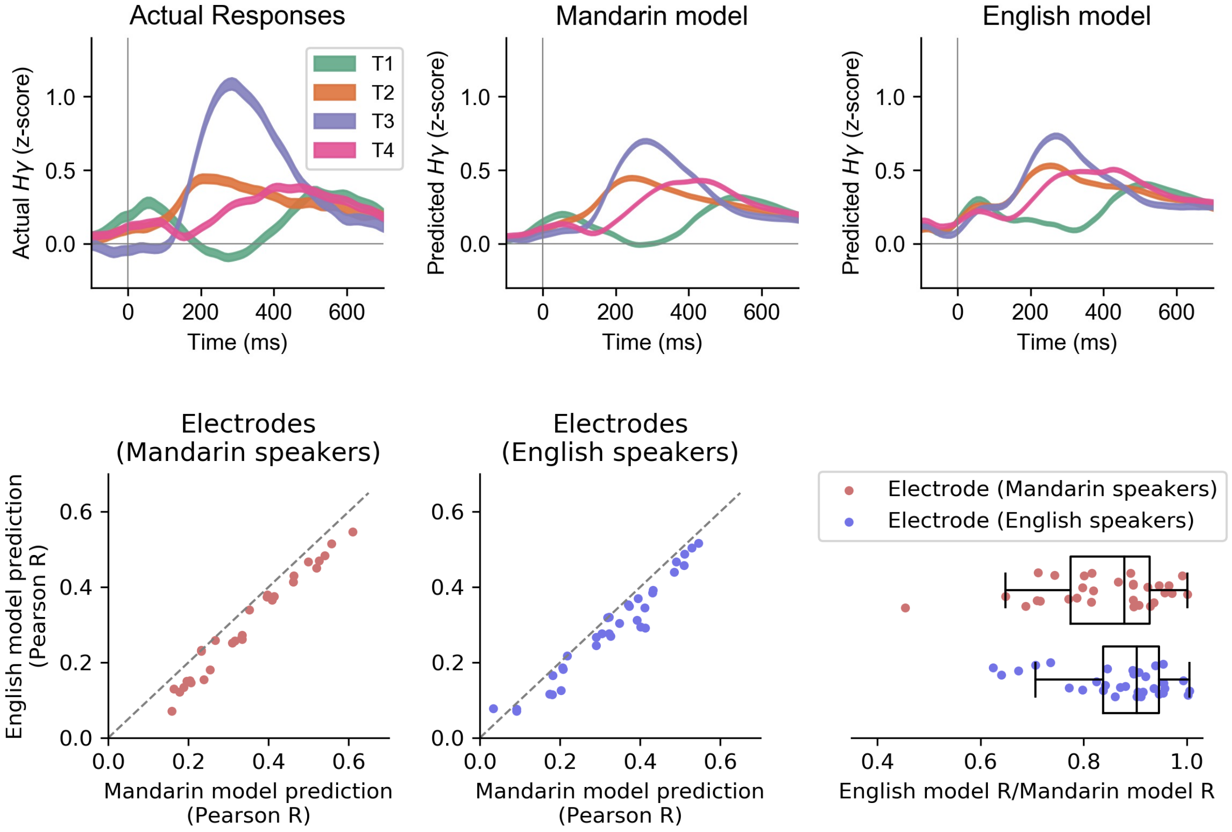
Same listener listen to speech in different languages

上海科技大学 ShanghaiTech University



Single electrode encoding of speaker-normalized pitch is language-independent

- Encoding model trained using English speech predicted neural response to lexical tones in Mandarin as good as Mandarin model.



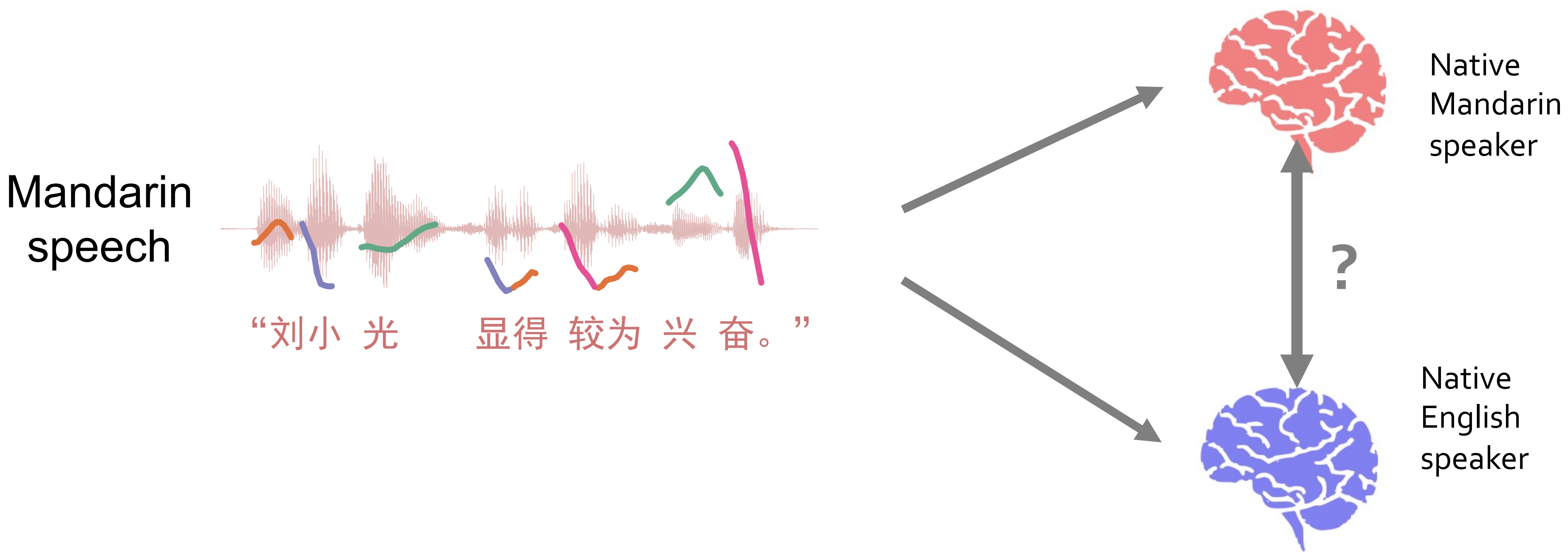
Research questions

上海科技大学 ShanghaiTech University

- What features are encoded in service of lexical tone representation?
  - Lower-level acoustic cues?
  - Complex intermediate features: speaker-normalized pitch (height and change)
  - Abstract tone category?

- Is the neural computation underlying lexical tone perception language-specific?
  - Single electrode encoding of speaker-normalized pitch is largely language-independent.
  - What about the STG population response?

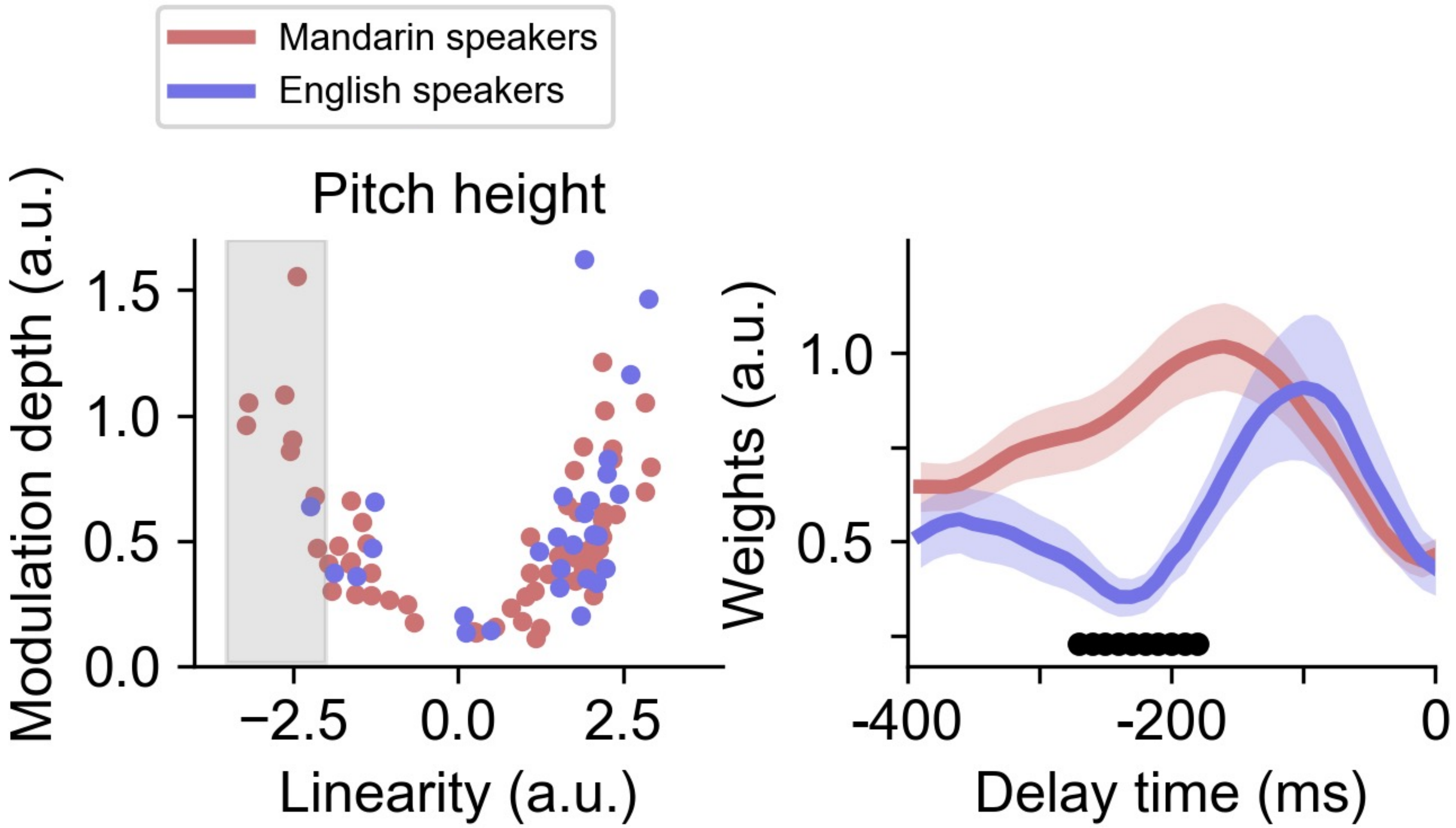
Different listeners listen to the same speech



Li et al., *Nat. Commun*, 2021

Mandarin speakers showed broader dynamic range and longer temporal integration window for pitch encoding in STG

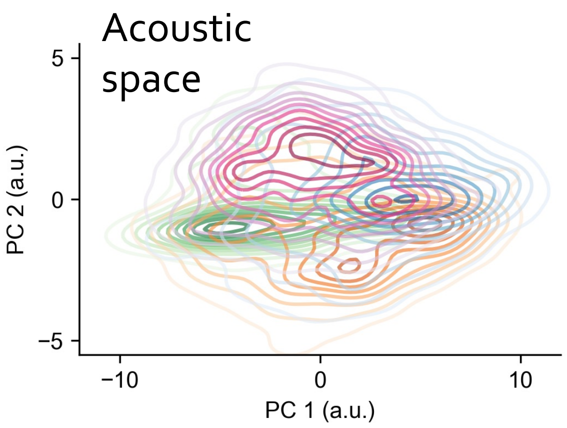
上海科技大学
ShanghaiTech University

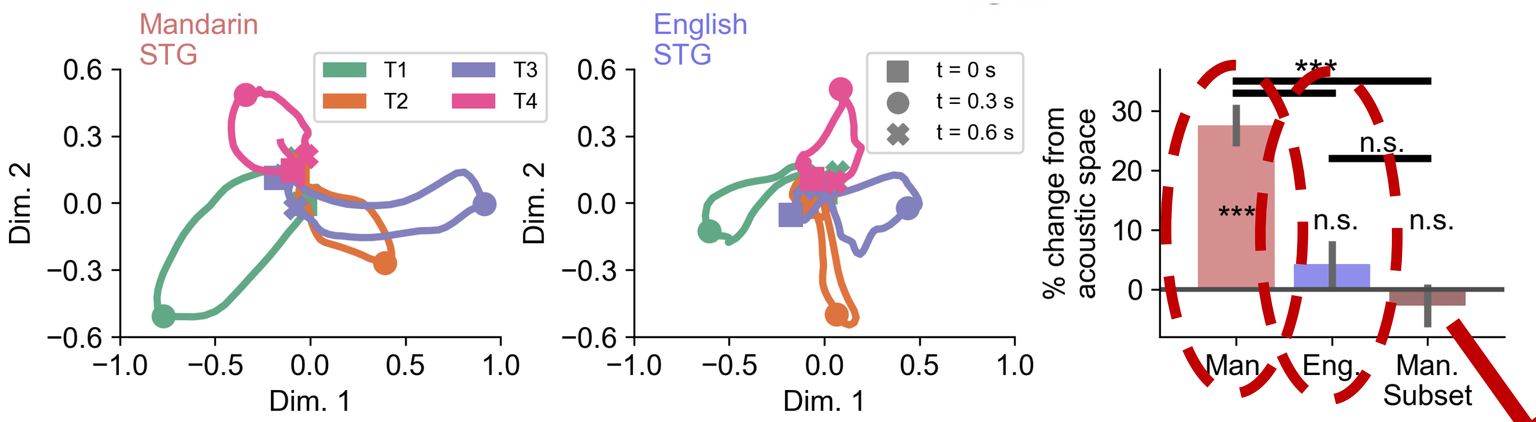


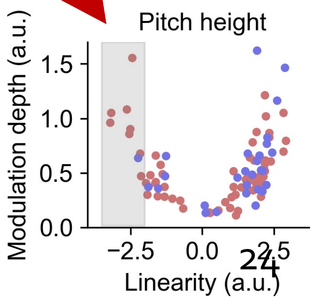
Li et al., *Nat. Commun*, 2021

Compare STG state space to acoustic space

- Tone decoding accuracy in STG population and acoustic space:
  - Mandarin speakers > Acoustic space = English speakers = Mandarin subset (take out negative coding electrodes)







Li et al., *Nat. Commun*, 2021

Research questions

- What features are encoded in service of lexical tone representation?
  - Lower-level acoustic cues?
  - Complex intermediate features: speaker-normalized pitch (height and change)
  - Abstract tone category?

- Is the neural computation underlying lexical tone perception language-specific?
  - Single electrode encoding of speaker-normalized pitch is largely language-independent.
  - What about the STG population response?

Research questions

上海科技大学
ShanghaiTech University

- What features are encoded in service of lexical tone representation?
  - Lower-level acoustic cues?
  - Complex intermediate features: speaker-normalized pitch (height and change)
  - Abstract tone category?

- Is the neural computation underlying lexical tone perception language-specific?
  - Single electrode encoding of speaker-normalized pitch is largely language-independent.
  - Population representation are influenced by language experience.

Marr's three levels of analysis

David Marr's three levels

- **Computational level**
  - cognitive task and problem
- **Algorithm level**
  - information representation and transformation
- **Implementation level**
  - implementation through interactions between basic elements

Neural coding

- Fit the underlying neural activity
- Verify theoretical hypotheses

e.g. relative pitch encoding, articulatory phonetic features, etc.

Computational model

Marr 1982

Marr’s three levels of analysis

David Marr’s three levels

- **Computational level**
  - cognitive task and problem
- **Algorithm level**
  - information representation and transformation
- **Implementation level**
  - implementation through interactions between basic elements

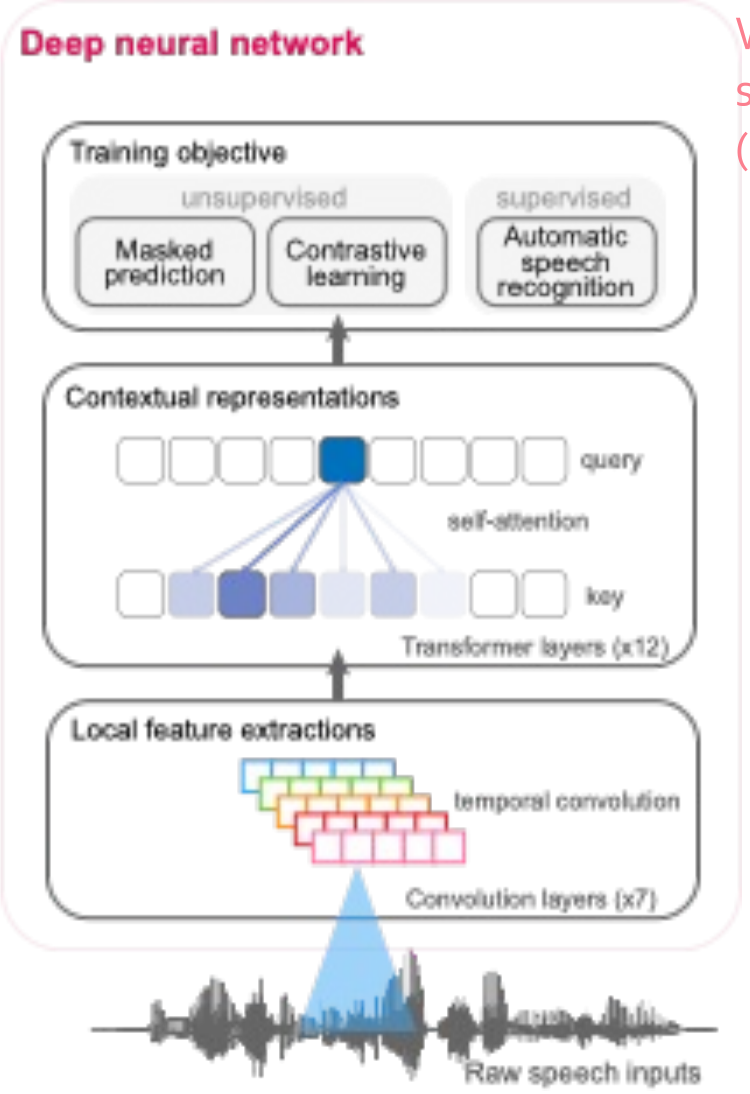
Neural coding

Computational model

- Fit the underlying neural activity
- Perform the cognitive tasks

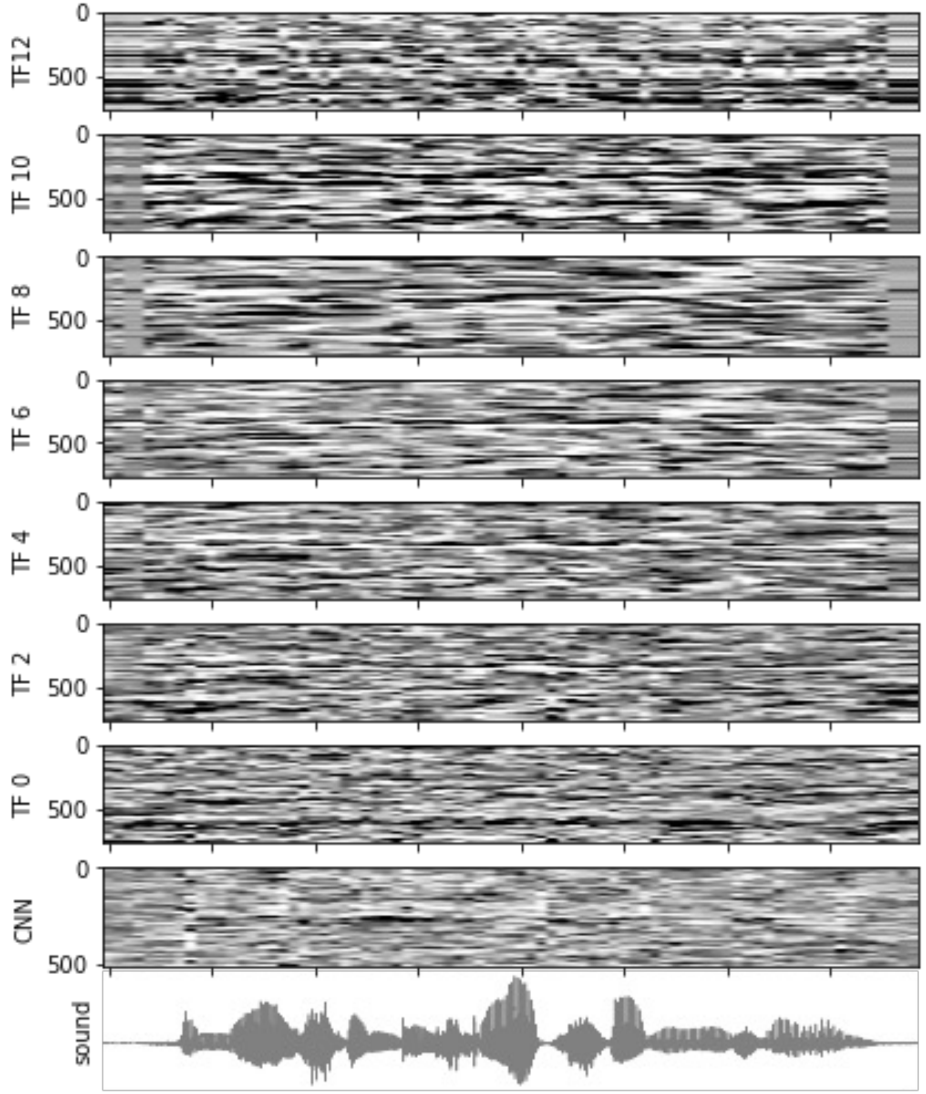
Marr 1982

State-of-the-art AI models for speech



Word error rate ~ 5% in speech recognition tasks (human ~4%)

internal representation sequences



Wav2Vec 2.0: Baevski et al. *NeurIPS* 2020;
HuBERT: Hsu et al. *ICASSP* 2021

Research questions

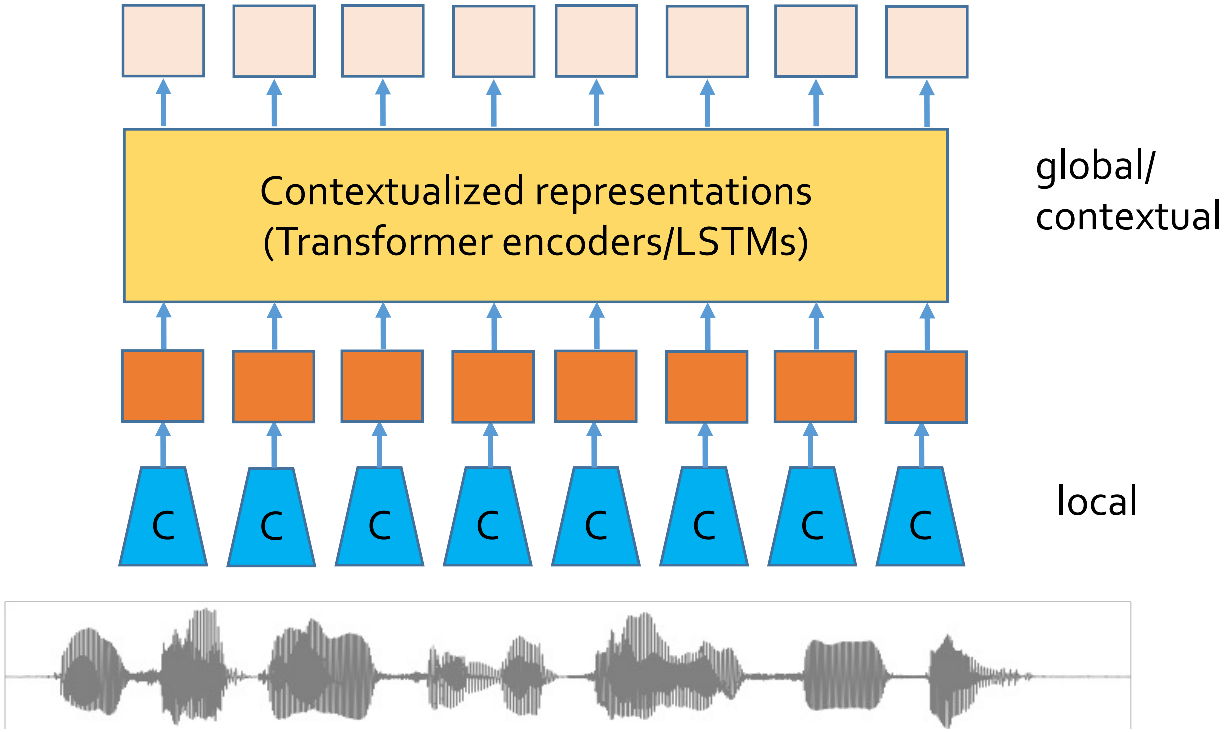
- What is a good deep neural network model for speech perception in auditory pathway?
  - Architecture: CNN-based models have been dominating
  - Training objective: supervised models have been dominating

- What are the key factors that make the DNN model good at predicting speech response in the brain?
  - Computations
  - Representations

Neural network models

- Same architecture w/ different training objectives
  - HuBERT (masked prediction)
  - Wav2Vec 2 unsupervised (contrastive learning)
  - Wav2Vec 2 supervised (ASR)
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
- Different architecture w/ same objectives
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
  - DeepSpeech 2 (ASR): LSTM

	Unsupervised objective	Supervised objective	Contextual units
W2V unsup	Contrastive learning	N/A	transformer
W2V sup	Contrastive learning	ASR	transformer
HuBERT	Masked prediction	N/A	transformer
HuBERT/W2V sup	N/A	ASR	transformer
DS2	N/A	ASR	LSTM



Neural network models

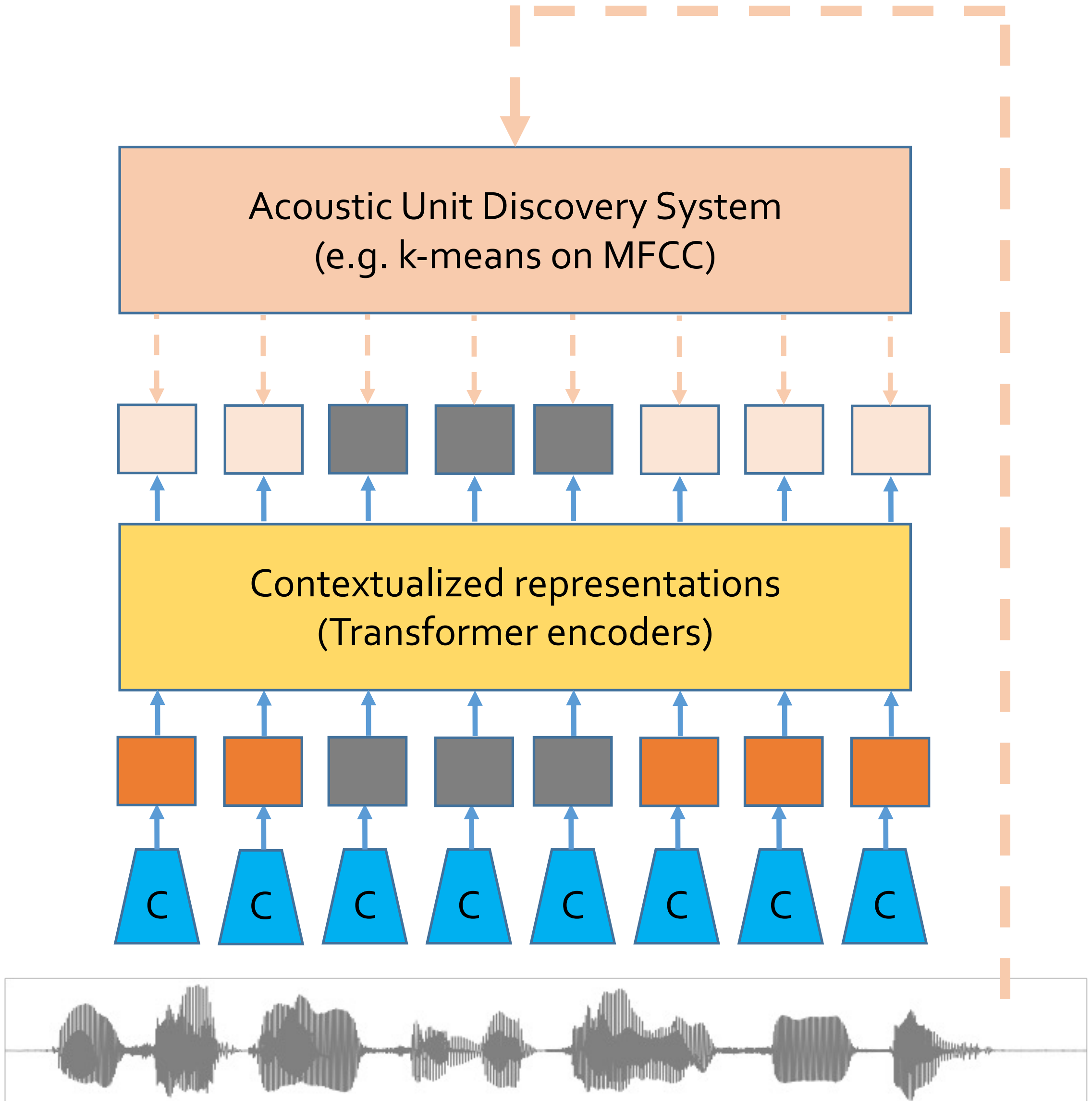
上海科技大学
ShanghaiTech University

- Same architecture w/ different training objectives
  - HuBERT (masked prediction)
  - Wav2Vec 2 unsupervised (contrastive learning)
  - Wav2Vec 2 supervised (ASR)
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
- Different architecture w/ same objectives
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
  - DeepSpeech 2 (ASR): LSTM

	Unsupervised objective	Supervised objective	Contextual units
W2V unsup	Contrastive learning	N/A	transformer
W2V sup	Contrastive learning	ASR	transformer
HuBERT	Masked prediction	N/A	transformer
HuBERT/W2V sup	N/A	ASR	transformer
DS2	N/A	ASR	LSTM

HuBERT

Hsu et al. 2021



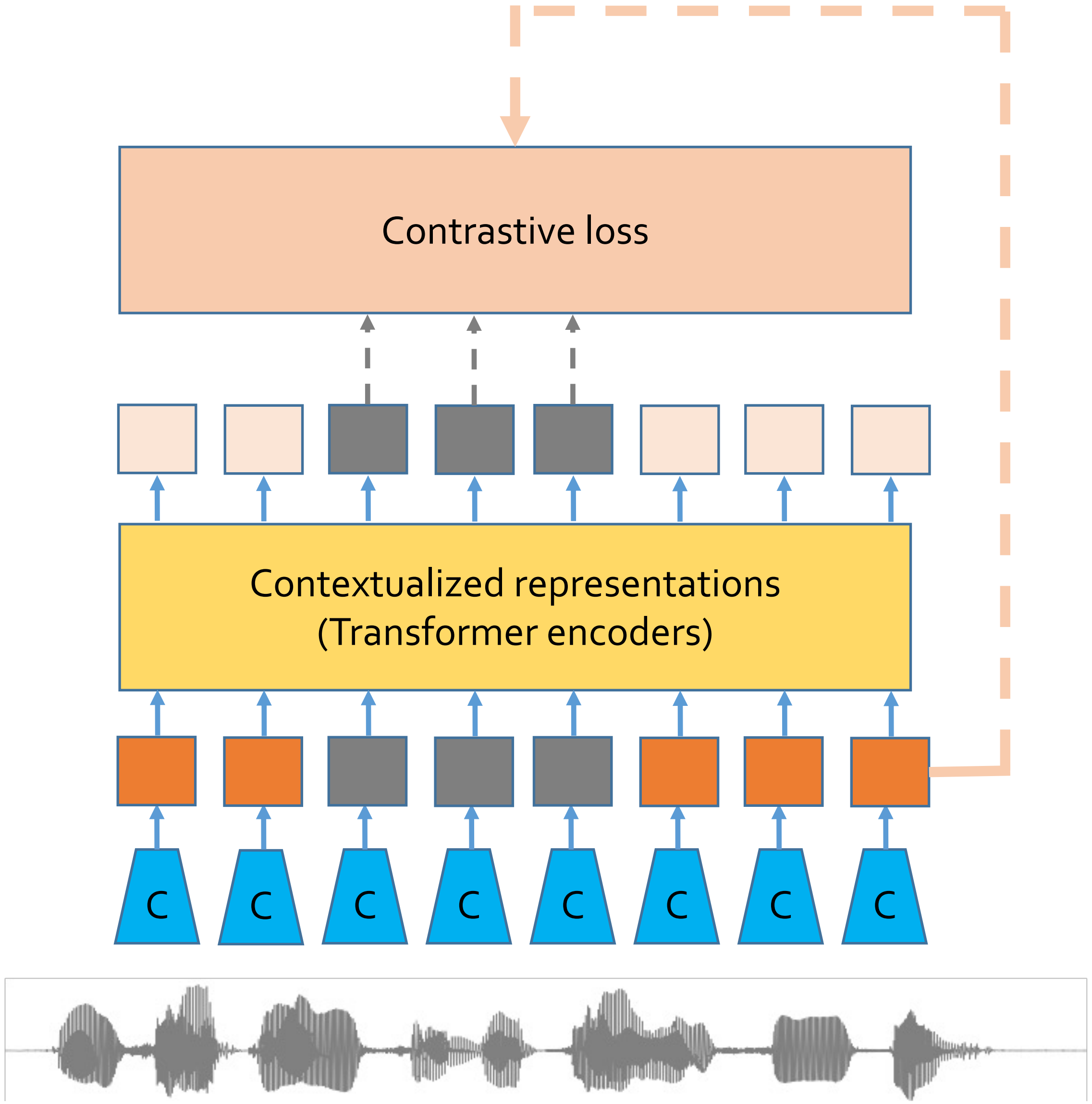
Neural network models

- Same architecture w/ different training objectives
  - HuBERT (masked prediction)
  - Wav2Vec 2 unsupervised (contrastive learning)
  - Wav2Vec 2 supervised (ASR)
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
- Different architecture w/ same objectives
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
  - DeepSpeech 2 (ASR): LSTM

	Unsupervised objective	Supervised objective	Contextual units
W2V unsup	Contrastive learning	N/A	transformer
W2V sup	Contrastive learning	ASR	transformer
HuBERT	Masked prediction	N/A	transformer
HuBERT/W2V sup	N/A	ASR	transformer
DS2	N/A	ASR	LSTM

Wav2Vec 2 unsupervised

Baevski et al. 2020



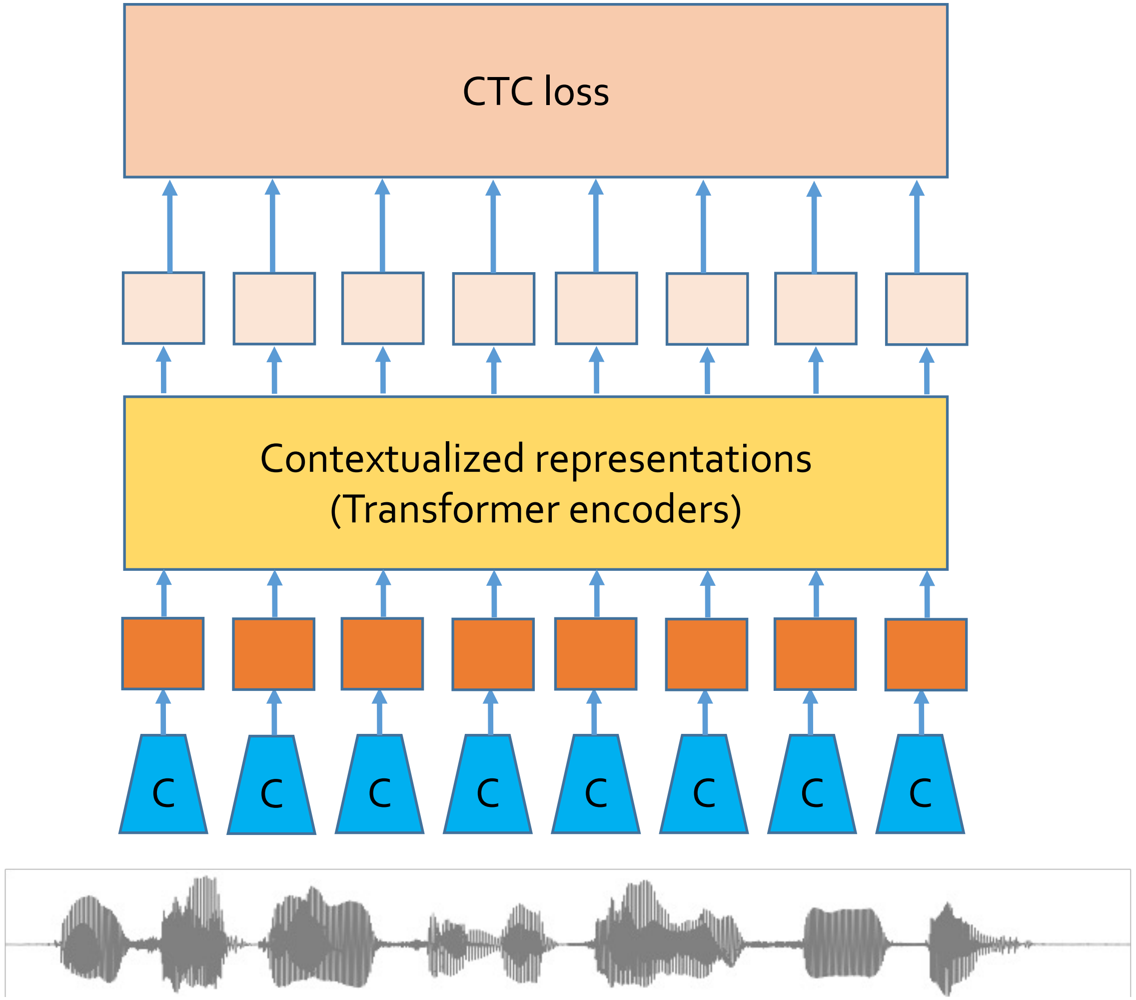
Neural network models

- Same architecture w/ different training objectives
  - HuBERT (masked prediction)
  - Wav2Vec 2 unsupervised (contrastive learning)
  - Wav2Vec 2 supervised (ASR)
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
- Different architecture w/ same objectives
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
  - DeepSpeech 2 (ASR): LSTM

	Unsupervised objective	Supervised objective	Contextual units
W2V unsup	Contrastive learning	N/A	transformer
W2V sup	Contrastive learning	ASR	transformer
HuBERT	Masked prediction	N/A	transformer
HuBERT/W2V sup	N/A	ASR	transformer
DS2	N/A	ASR	LSTM

Wav2Vec 2 supervised

Baevski et al. 2020



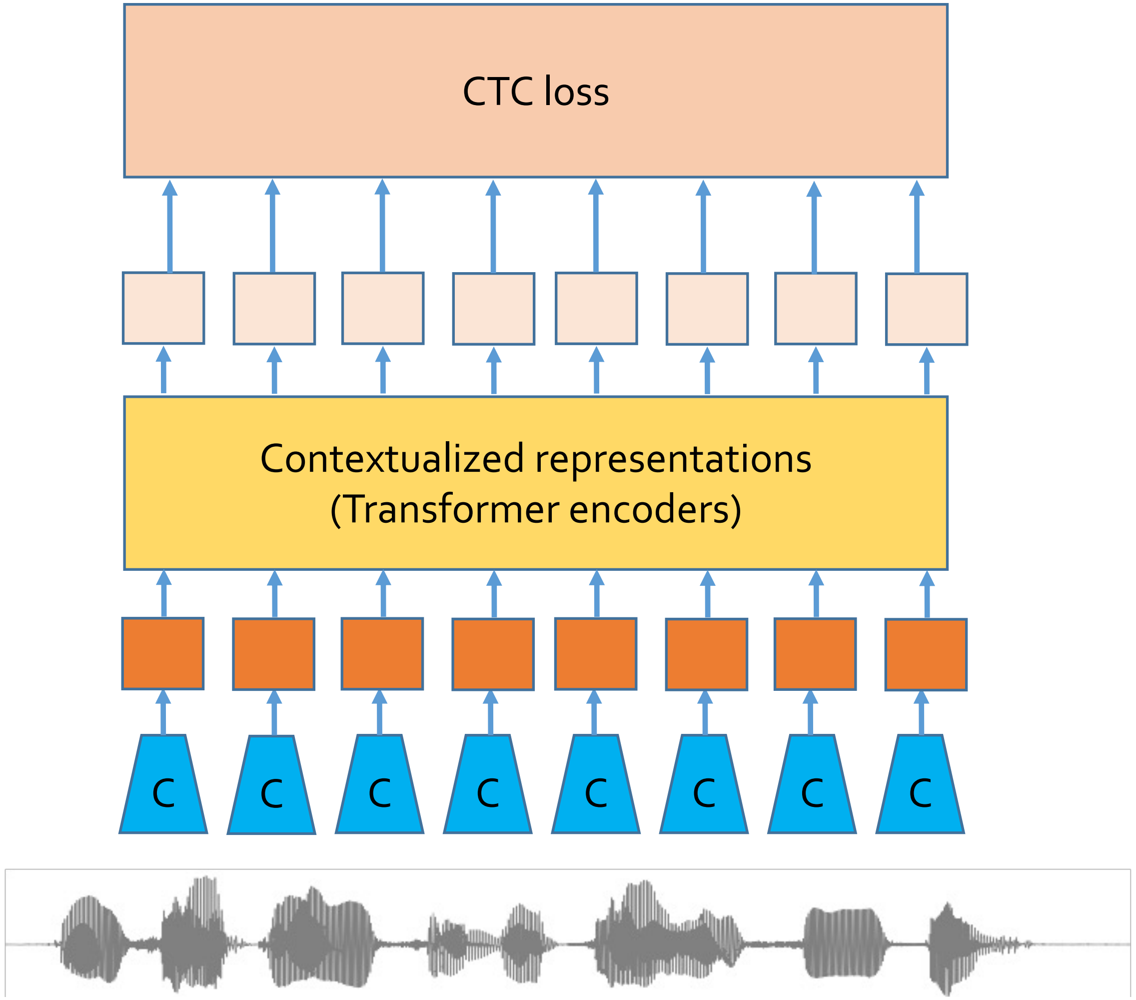
Neural network models

- Same architecture w/ different training objectives
  - HuBERT (masked prediction)
  - Wav2Vec 2 unsupervised (contrastive learning)
  - Wav2Vec 2 supervised (ASR)
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
- Different architecture w/ same objectives
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
  - DeepSpeech 2 (ASR): LSTM

	Unsupervised objective	Supervised objective	Contextual units
W2V unsup	Contrastive learning	N/A	transformer
W2V sup	Contrastive learning	ASR	transformer
HuBERT	Masked prediction	N/A	transformer
HuBERT/W2V sup	N/A	ASR	transformer
DS2	N/A	ASR	LSTM

HuBERT/Wav2Vec 2
Pure supervised

Baevski et al. 2020



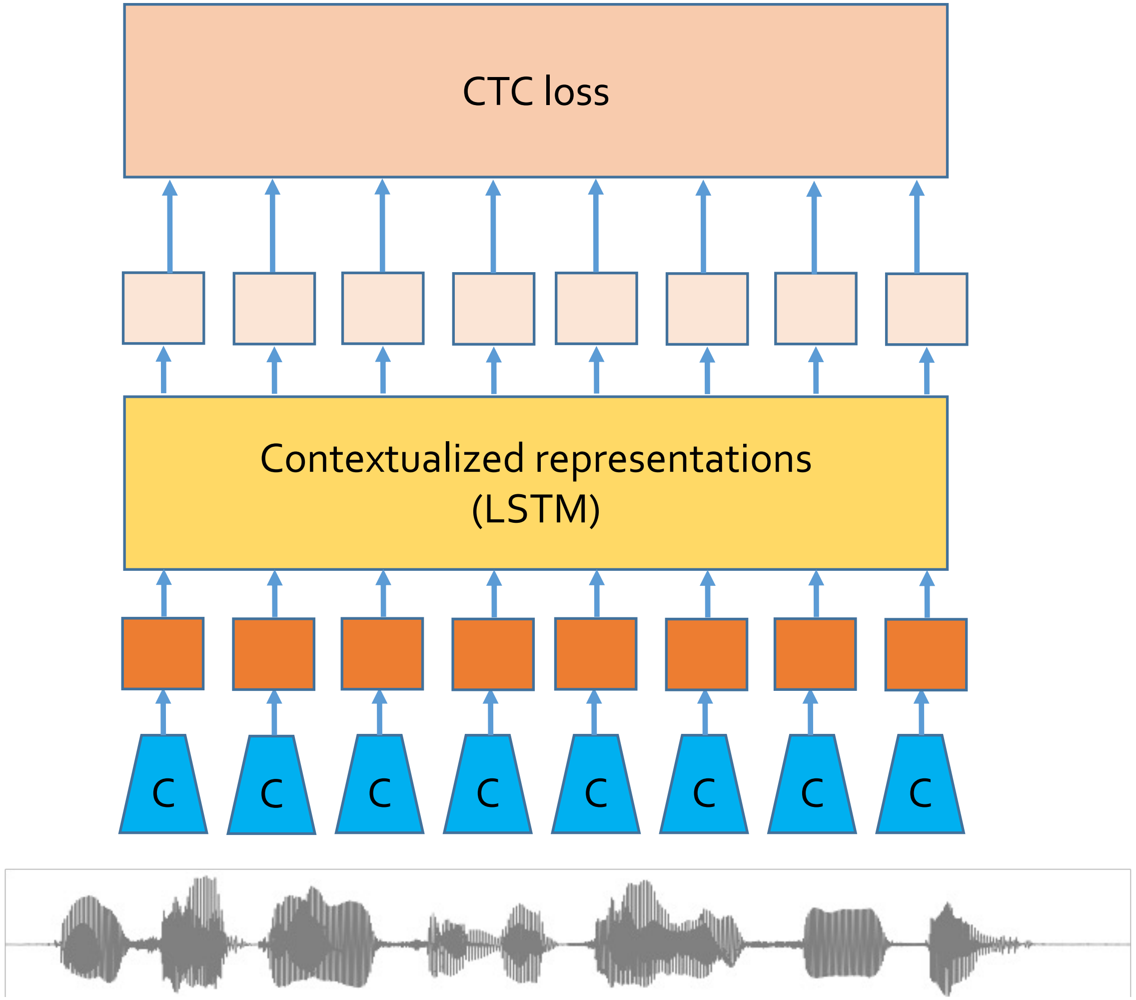
Neural network models

- Same architecture w/ different training objectives
  - HuBERT (masked prediction)
  - Wav2Vec 2 unsupervised (contrastive learning)
  - Wav2Vec 2 supervised (ASR)
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
- Different architecture w/ same objectives
  - HuBERT/Wav2Vec 2 pure supervised (ASR)
  - DeepSpeech 2 (ASR): LSTM

	Unsupervised objective	Supervised objective	Contextual units
W2V unsup	Contrastive learning	N/A	transformer
W2V sup	Contrastive learning	ASR	transformer
HuBERT	Masked prediction	N/A	transformer
HuBERT/W2V sup	N/A	ASR	transformer
DS2	N/A	ASR	LSTM

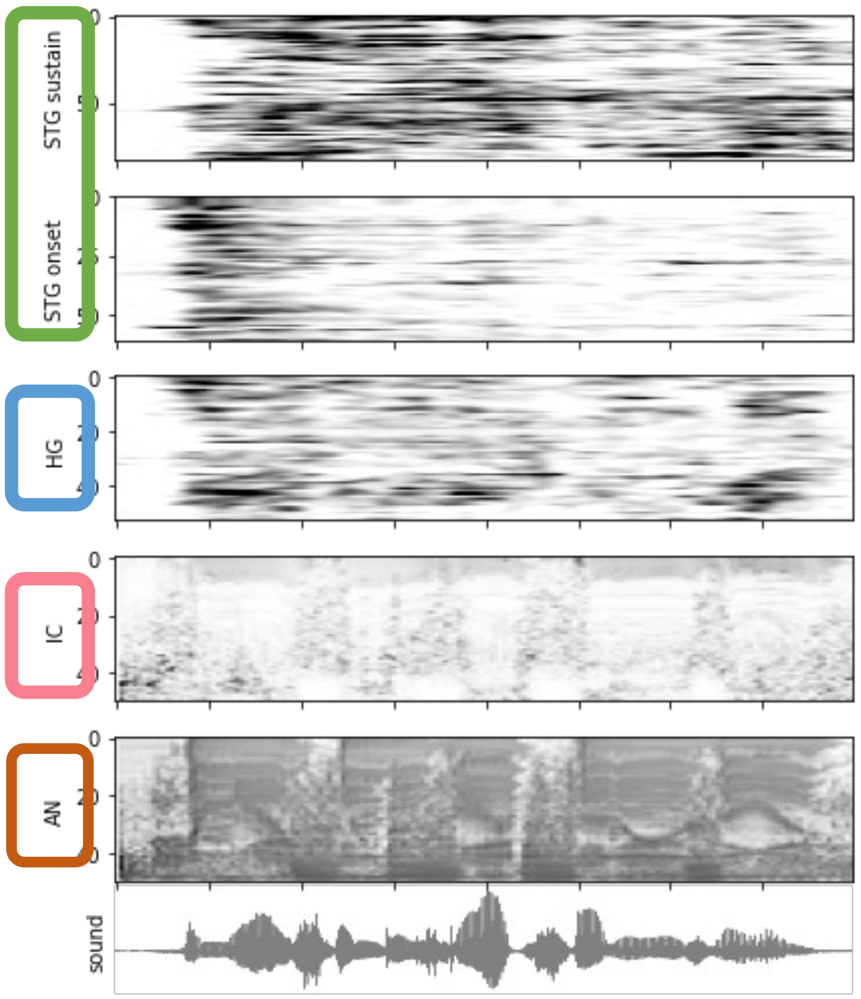
DeepSpeech 2

Amodei et al. 2016



Transformation of speech sound into phonetic units

Neural responses



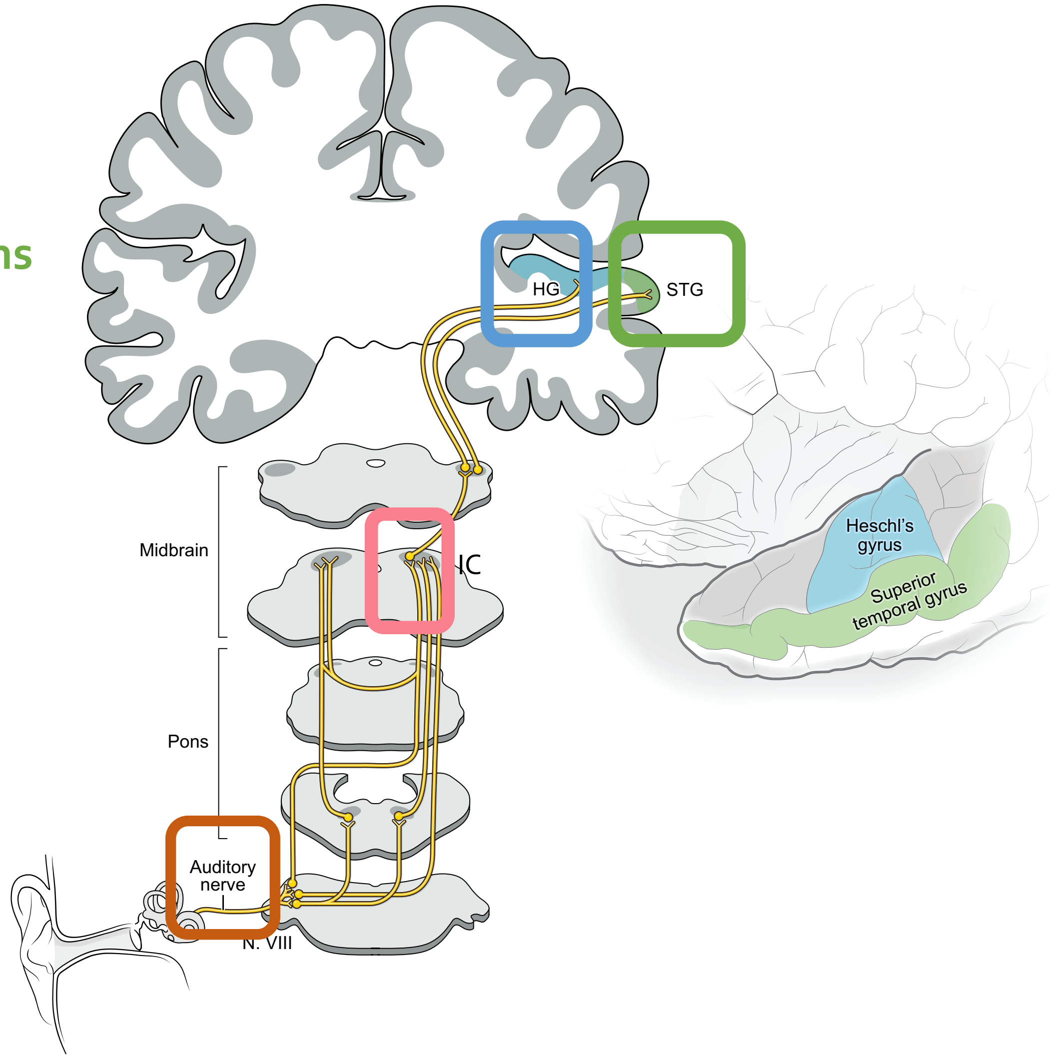
He moistened his lips uneasily.

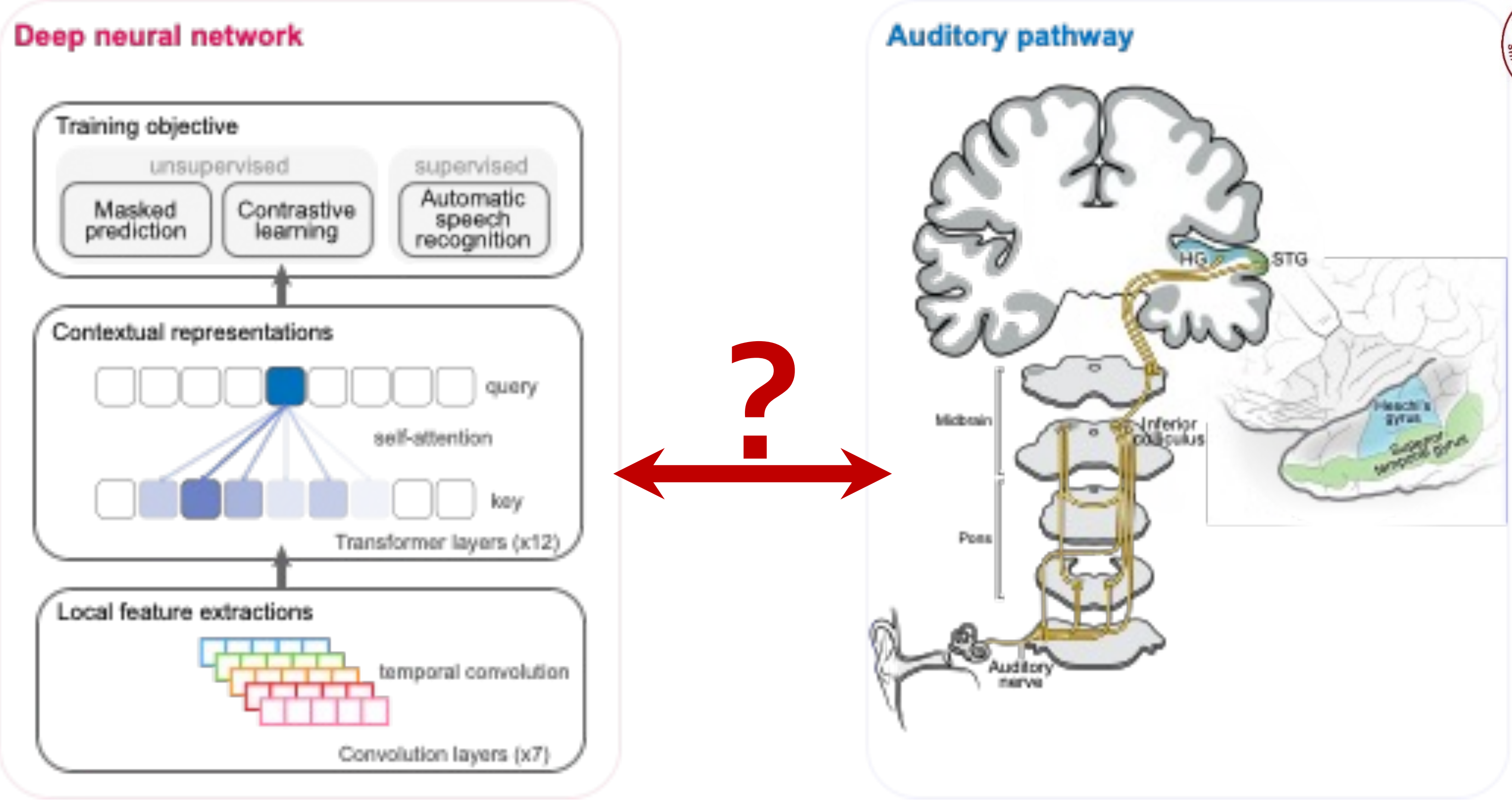
Complex time-frequency patterns & extended dynamics

Narrow band frequency tuning

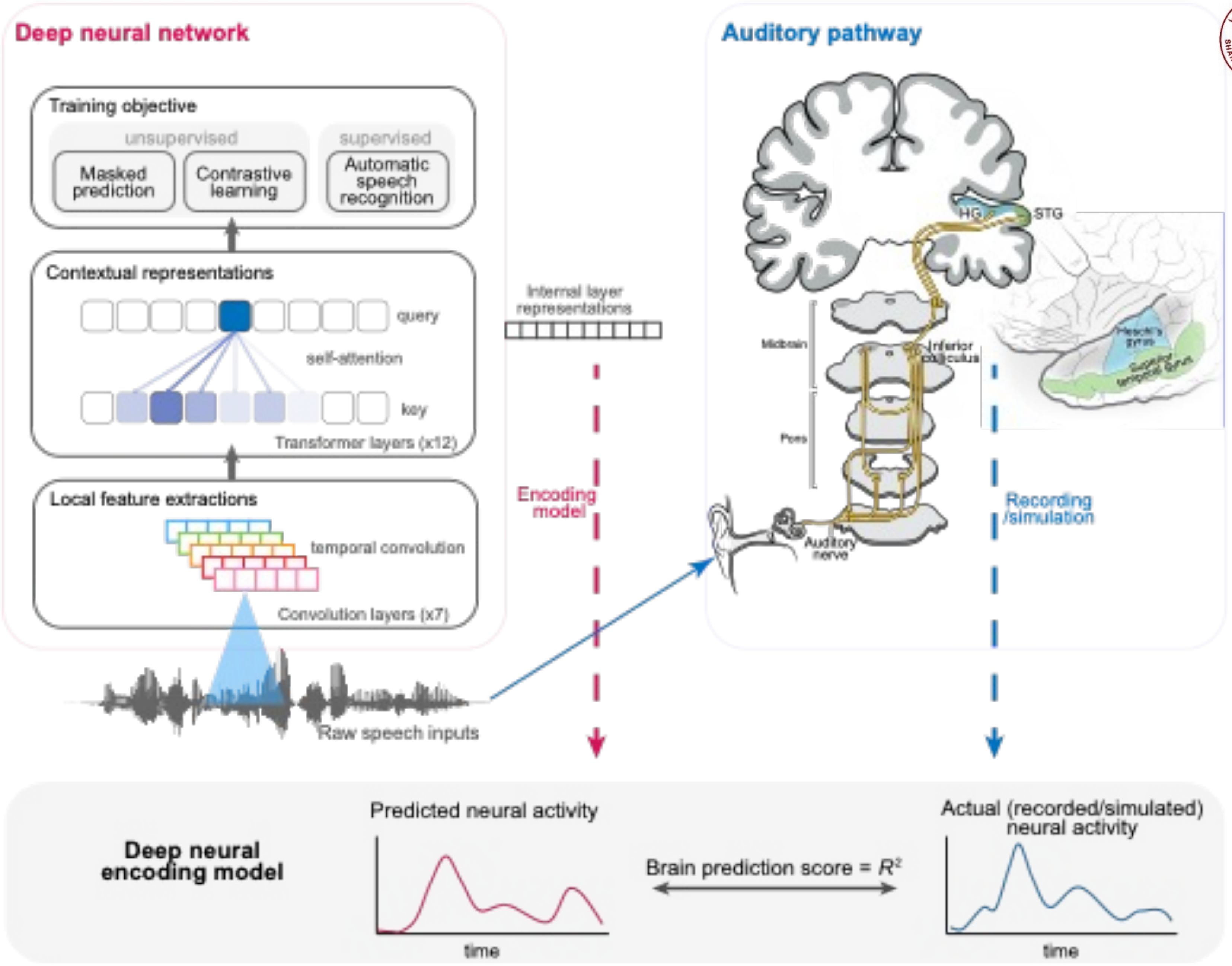
Band-pass and band-reject

Frequency selectivity




Deep neural network
Training objective
unsupervised
supervised
Masked prediction
Contrastive learning
Automatic speech recognition
Contextual representations
query
self-attention
key
Transformer layers (x12)
Local feature extractions
temporal convolution
Convolution layers (x7)
?
Auditory pathway
HG
STG
Midbrain
Pons
Inferior colliculus
Auditory nerve

上海科技大学 ShanghaiTech University

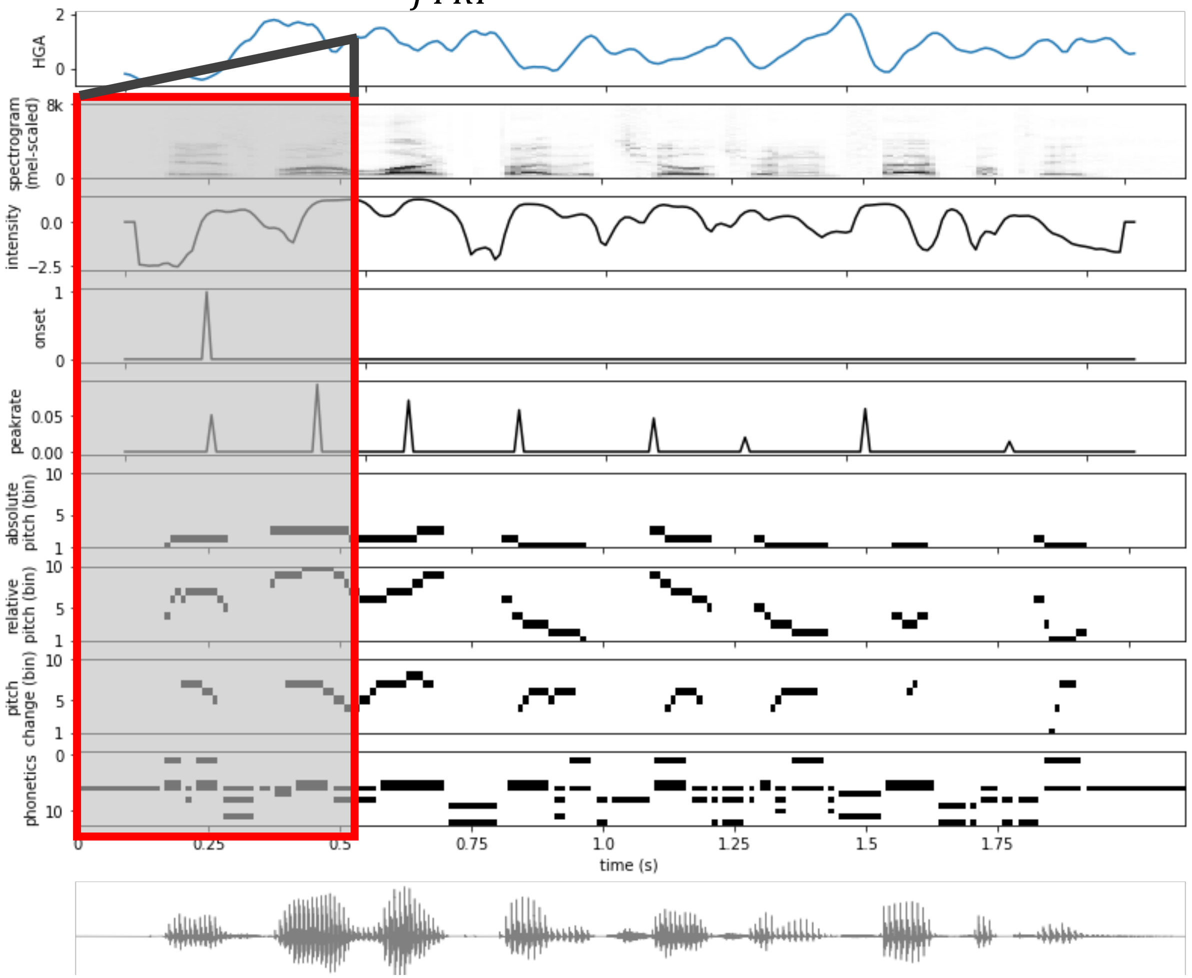


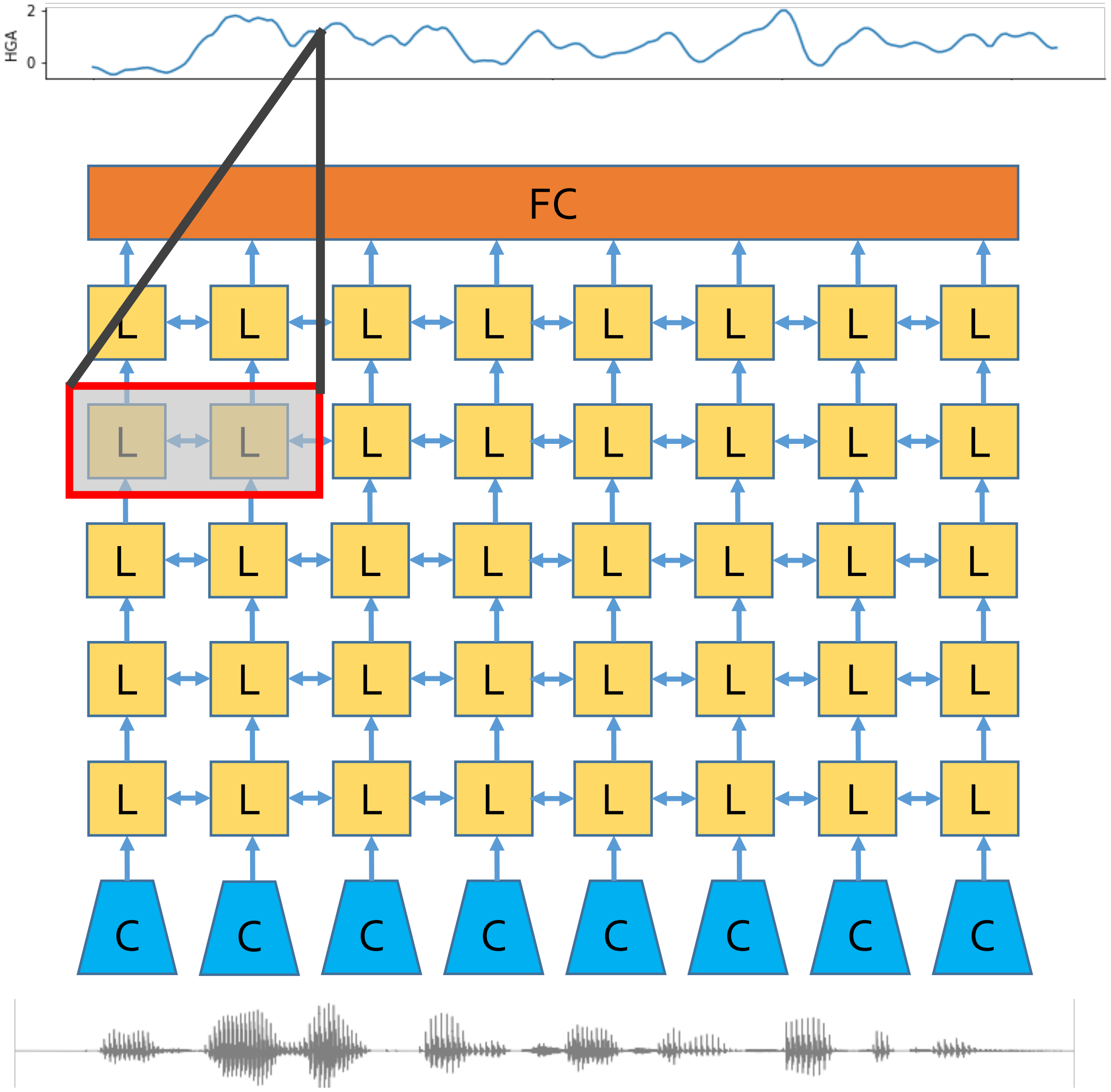
Comparing encoding models

R^2_{fTRF}

$$y_{predict} = X_{feature}\beta$$

R^2_{NN}



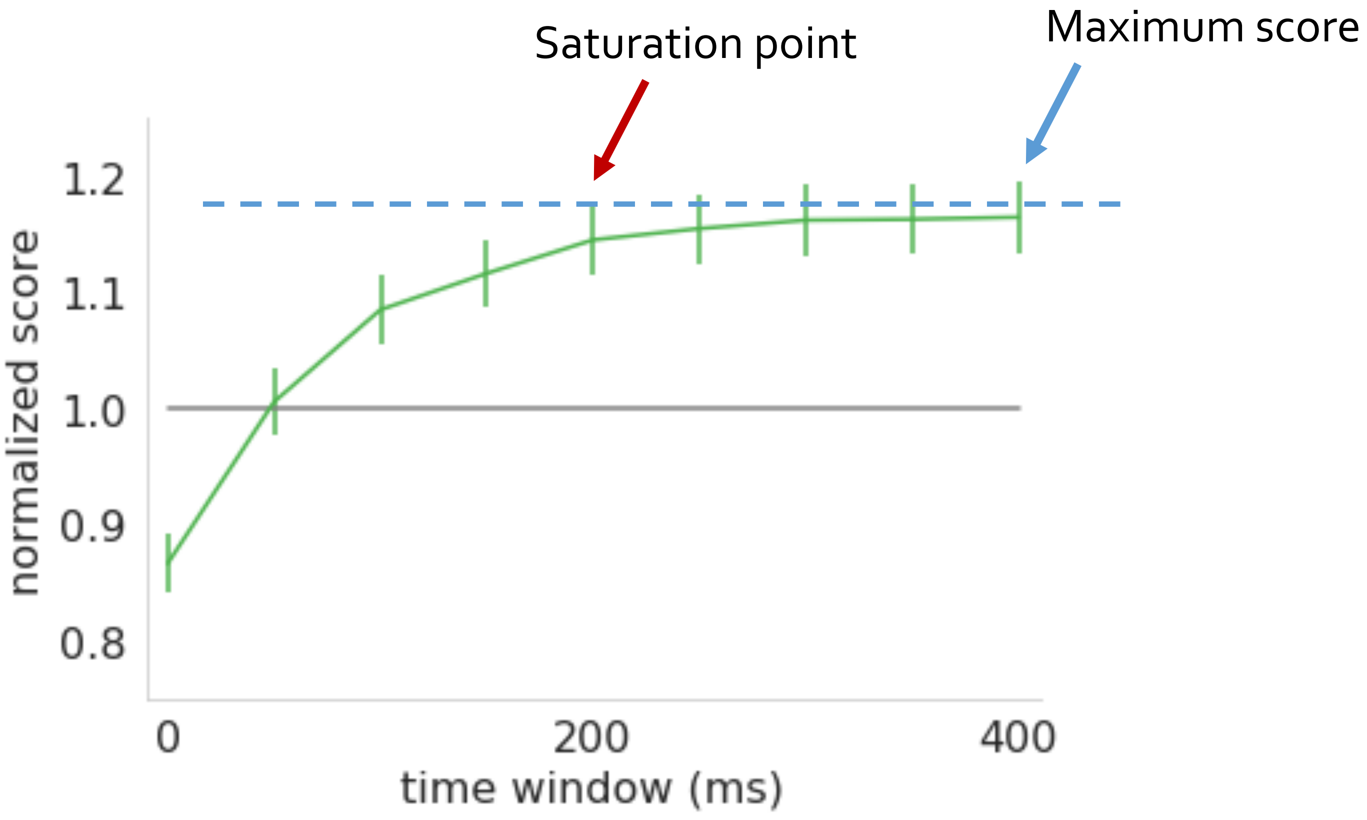


Brain prediction score

$$R^2_{normed} = \frac{R^2_{NN}}{R^2_{fTRF}}$$

Encoding models

- Metrics that quantify the performance of different encoding models
  - **Maximum prediction score**: maximum over all time window lengths
  - **Saturation point**: the minimum time window length such that maximum score is within mean + 1 s.e.m.



Research questions

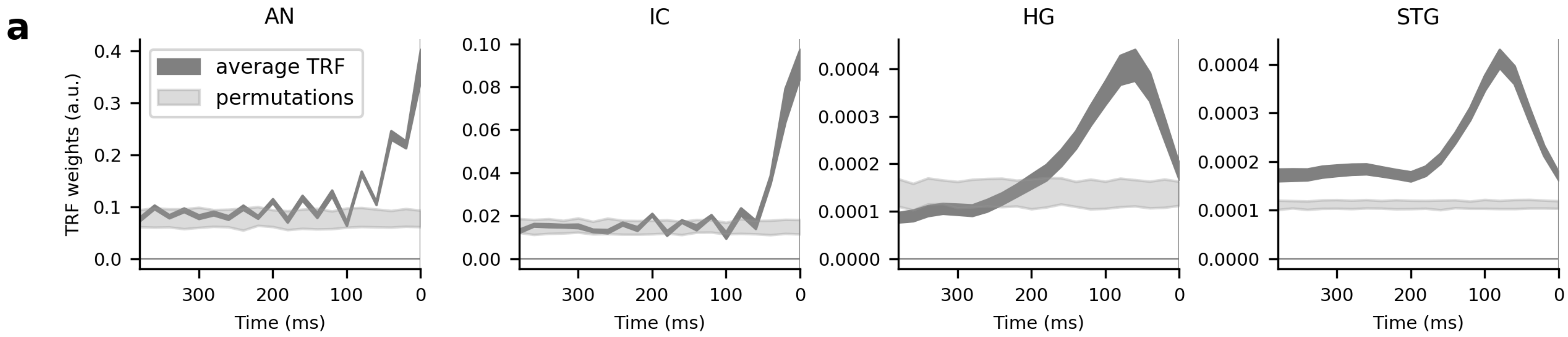
上海科技大学 ShanghaiTech University

- What is a good deep neural network model for speech perception in auditory pathway?
  - Architecture: CNN-based models have been dominating
  - Training objective: supervised models have been dominating

- What are the key factors that make the DNN model good at predicting speech response in the brain?
  - Computations:
  - Representations

What's the best model for each area?

- Different areas have drastically different temporal response profiles

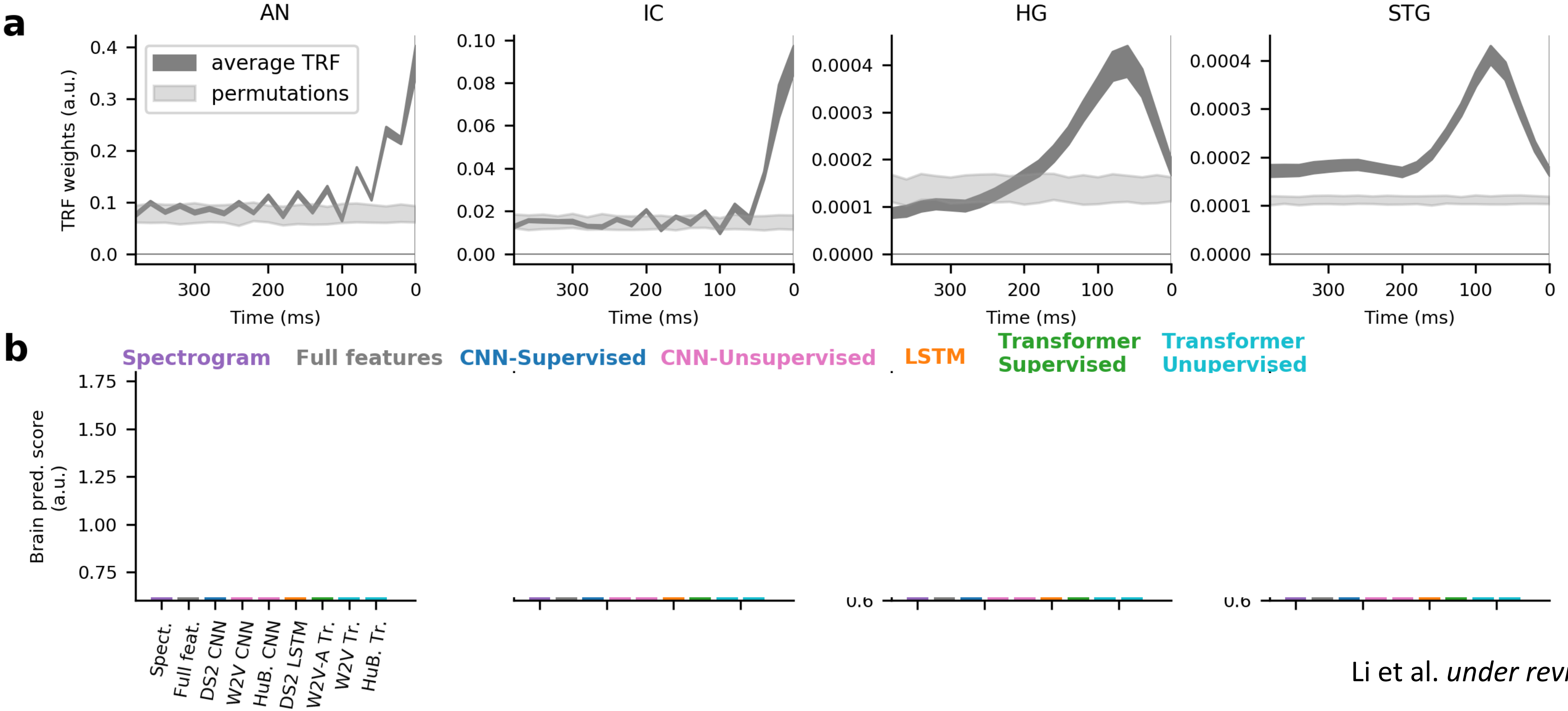


Li et al. *under review*

What’s the best model for each area?

上海科技大学 ShanghaiTech University

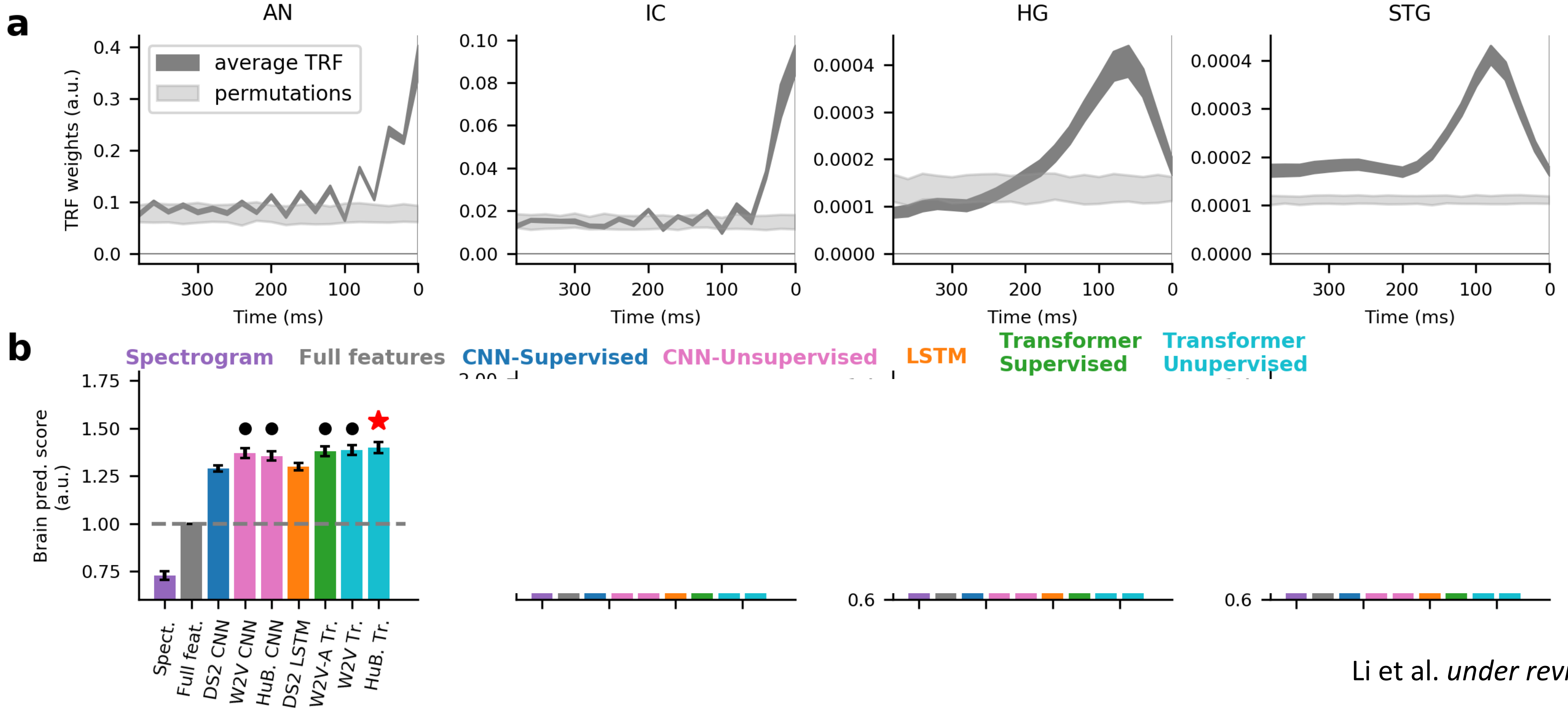
- Static nonlinear filters (CNN) is good for AN



Li et al. *under review*

What's the best model for each area?

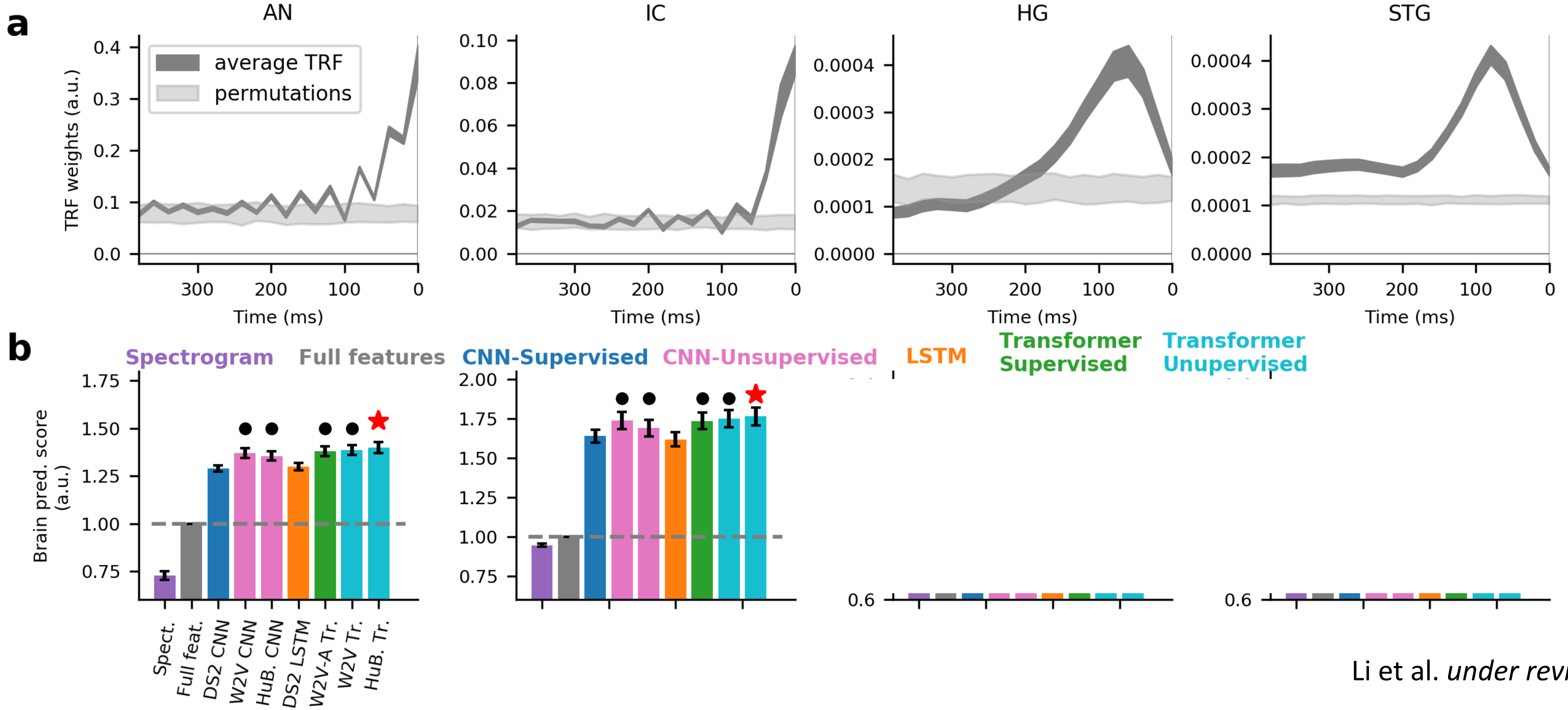
- Static nonlinear filters (CNN) is good for AN



Li et al. *under review*

What's the best model for each area?

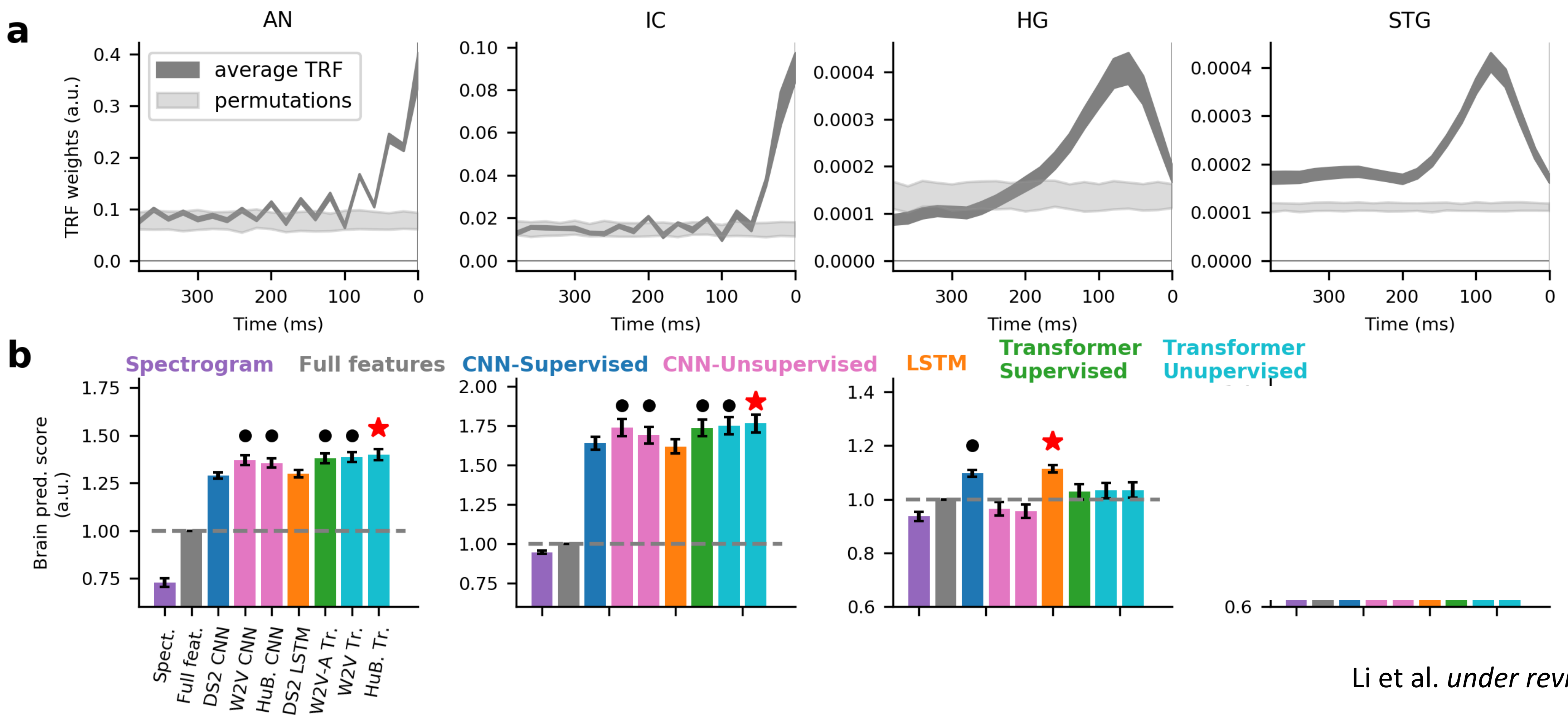
- Static nonlinear filters (CNN) is good for AN, IC



Li et al. *under review*

What’s the best model for each area?

- Static nonlinear filters (CNN) is good for AN, IC & HG

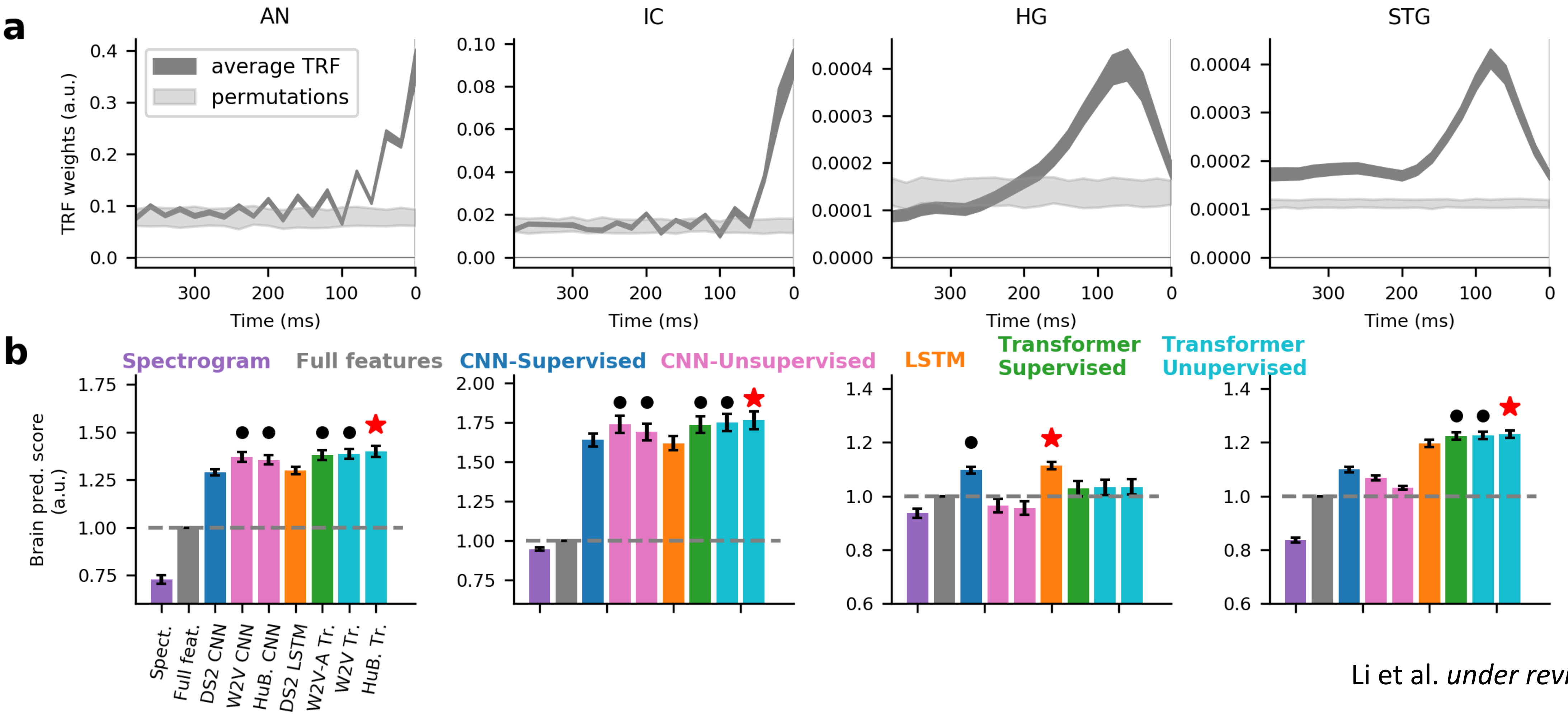


Li et al. *under review*

What's the best model for each area?

上海科技大学
ShanghaiTech University

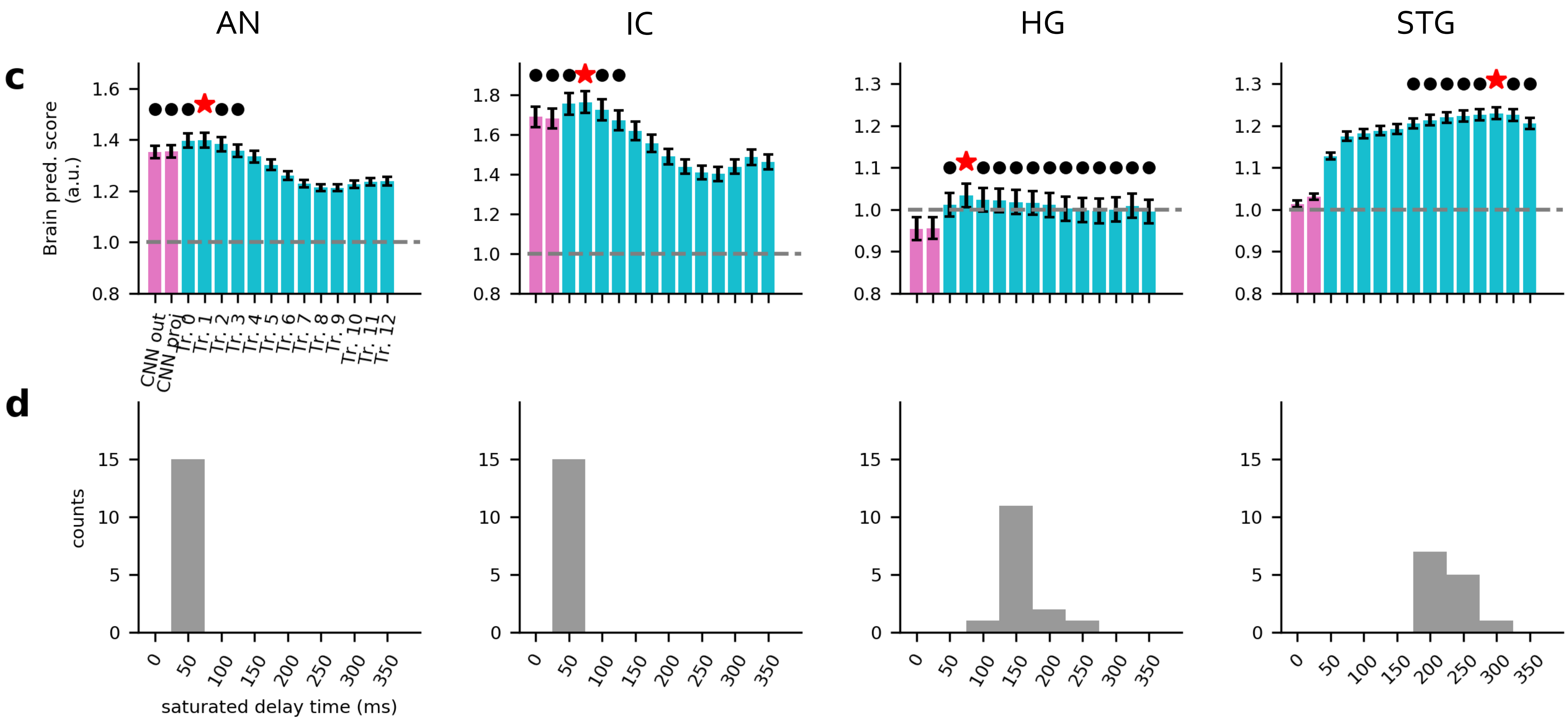
- Static nonlinear filters (CNN) is good for AN, IC & HG
- Contextual models (LSTM & Transformer) outperforming CNN & feature models in STG
- Unsupervised models perform as good as supervised models, if not better



Li et al. *under review*

The early to later layers in the same deep neural networks trained to learn speech representations correlate to the AN-Midbrain-STG pathway

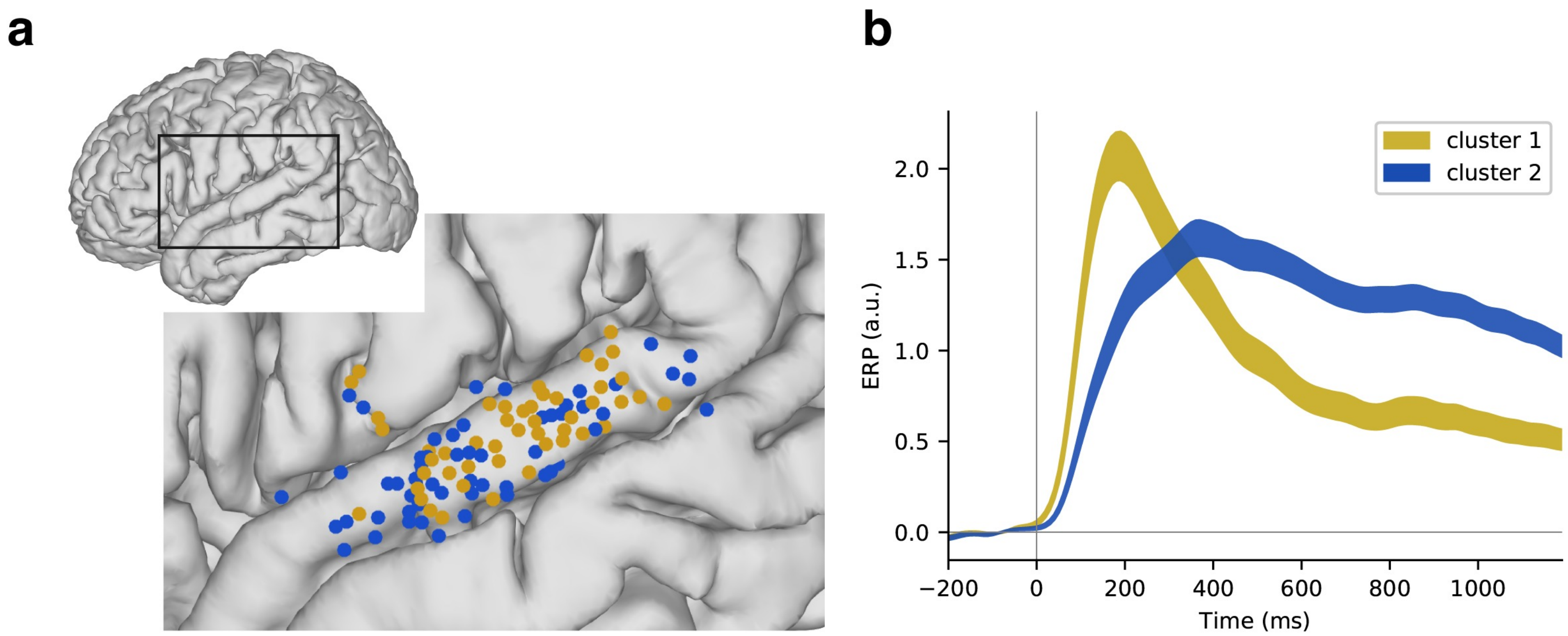
- Hierarchy within the same unsupervised model (HuBERT)



Li et al. *under review*

Clustering STG electrodes according to response profiles

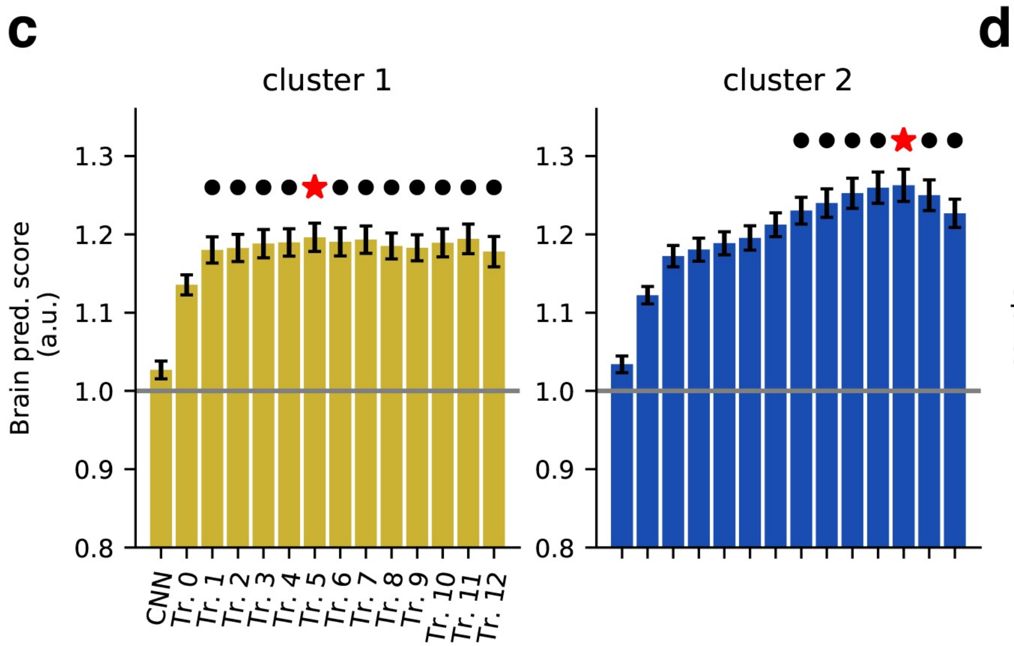
- NMF and clustering into onset and sustained populations

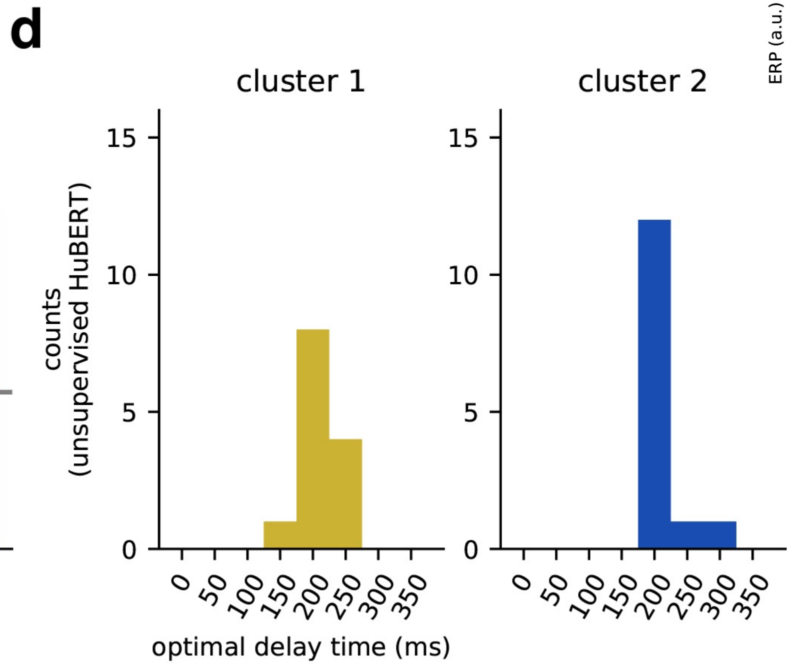


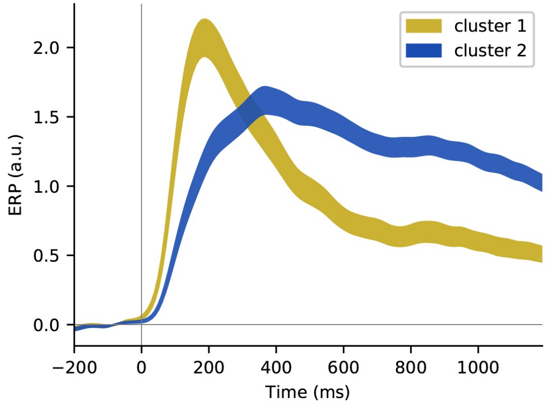
Li et al. *under review*

Functional subpopulations in STG correlate to different contextual representation layers in DNN

- DNN maintains the transient onset representation throughout the processing pipeline
- Later layers represent both transient and sustained representations in parallel







Li et al. *under review*

Research questions

上海科技大学 ShanghaiTech University

- What is a good deep neural network model for speech perception in auditory pathway?

- What are the key factors that make the DNN model good at predicting speech response in the brain?

Research questions

上海科技大学 ShanghaiTech University

- What is a good deep neural network model for speech perception in auditory pathway?

  - The early to later layers in the deep neural networks trained to learn speech representations correlate to the ascending AN-Midbrain-STG auditory pathway
  - Functional subpopulations in STG correlate to different contextual representation layers in DNN
  - The general results are consistent across network architecture and training objectives

- What are the key factors that make the DNN model good at predicting speech response in the brain?

Research questions

上海科技大学 ShanghaiTech University

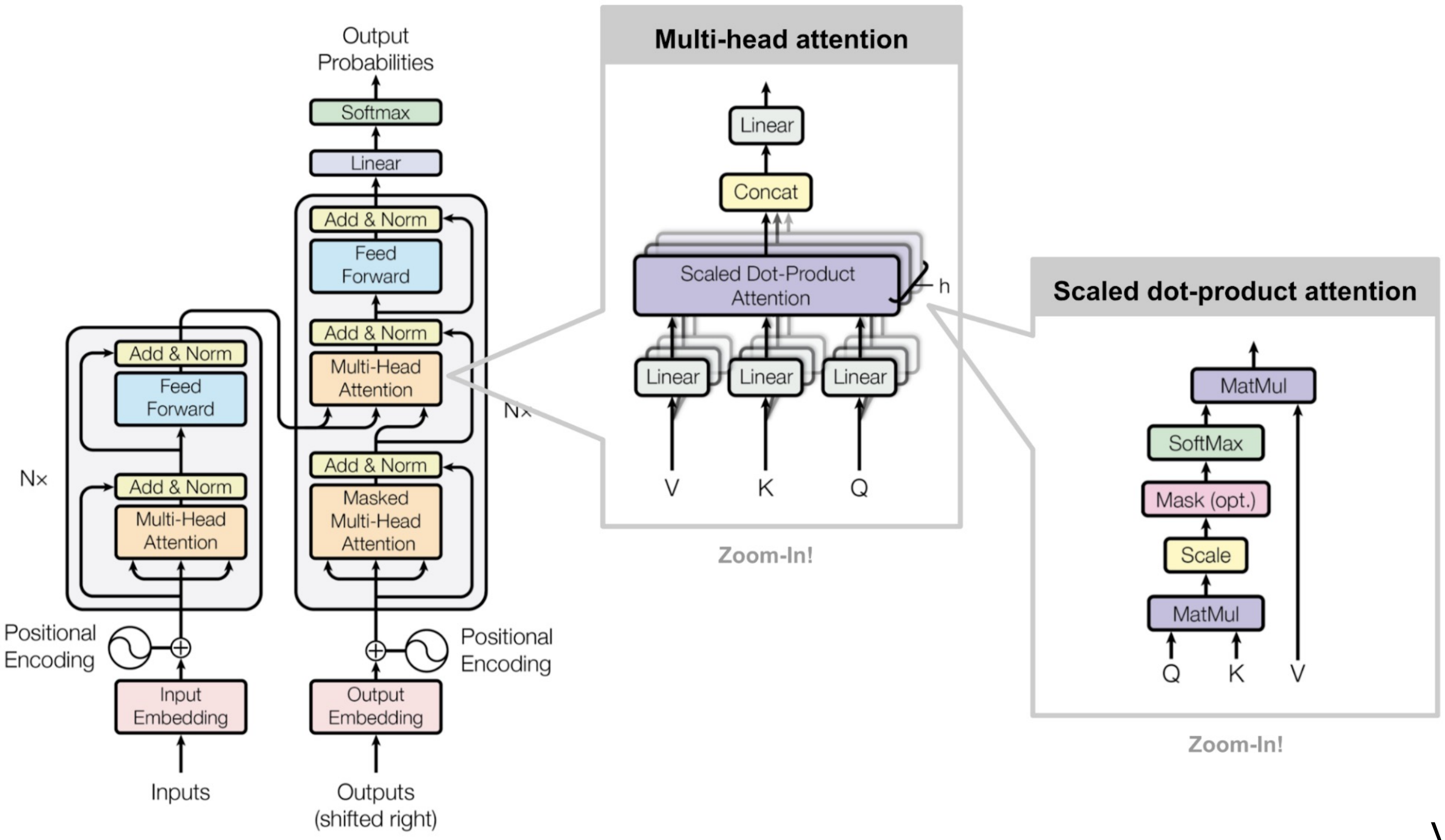
- What is a good deep neural network model for speech perception in auditory pathway?
    - The early to later layers in the deep neural networks trained to learn speech representations correlate to the ascending AN-Midbrain-STG auditory pathway
    - Functional subpopulations in STG correlate to different contextual representation layers in DNN
    - The general results are consistent across network architecture and training objectives

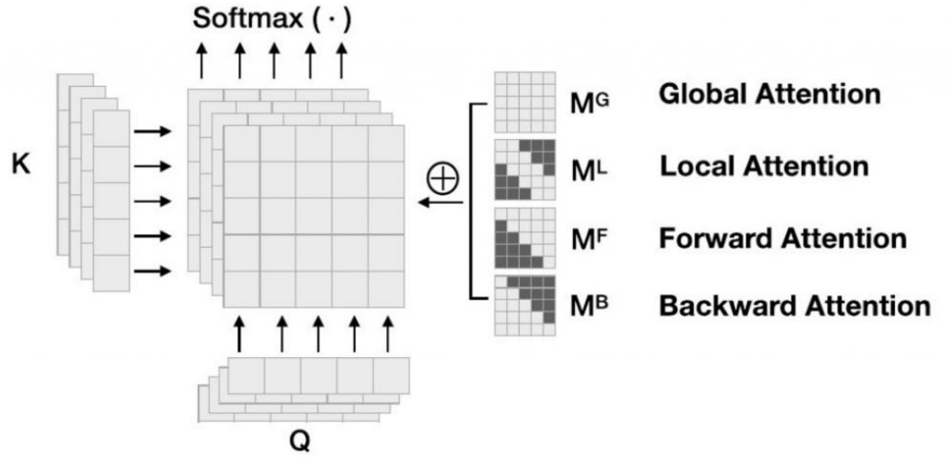
- What are the key factors that make the DNN model good at predicting speech response in the brain?

Context dependent computations in Transformer encoders

上海科技大学 ShanghaiTech University

- Transformer uses self-attention to extract context dependent information dynamically

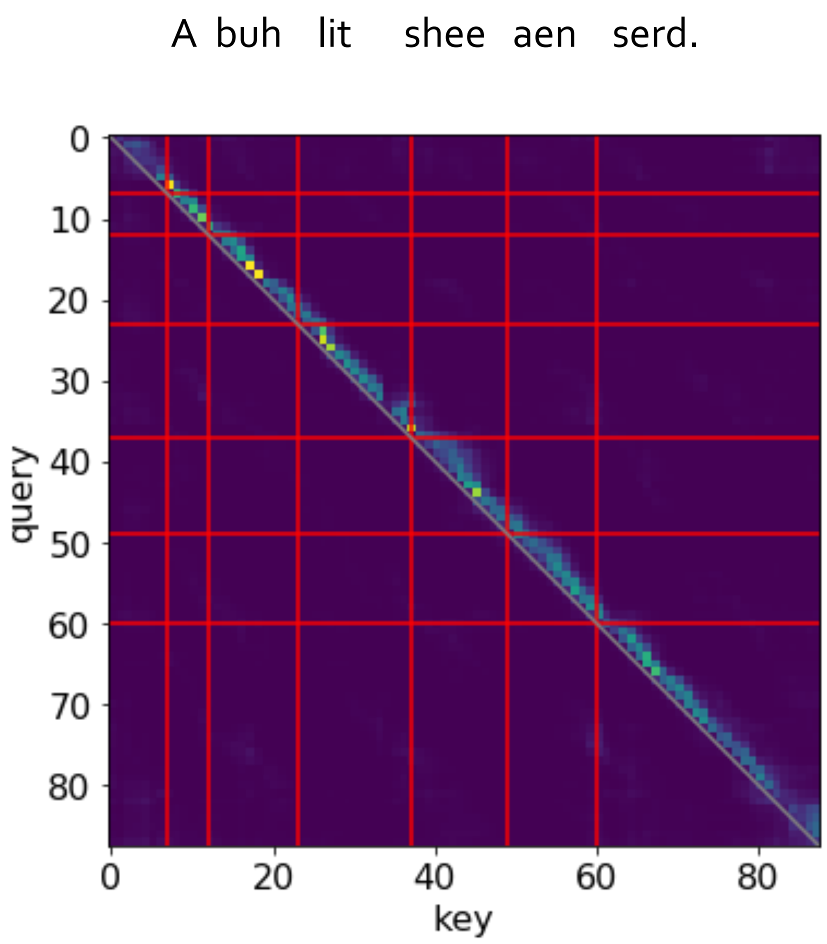




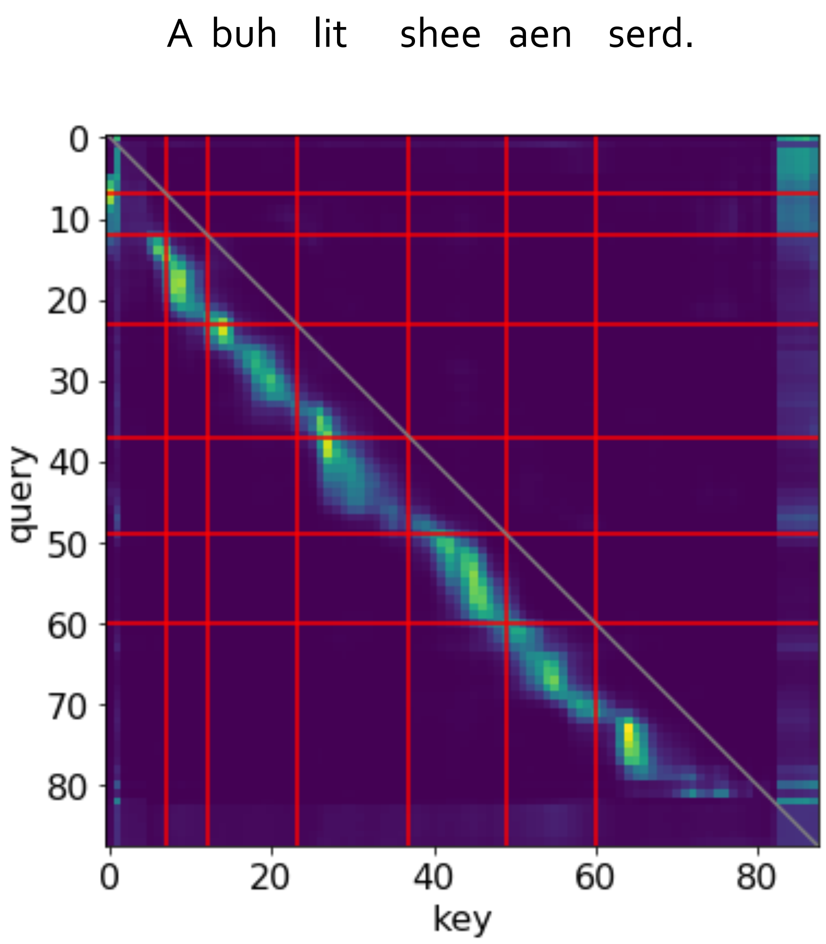
Vaswani et al. *NeurIPS* 2017

Context dependent computations

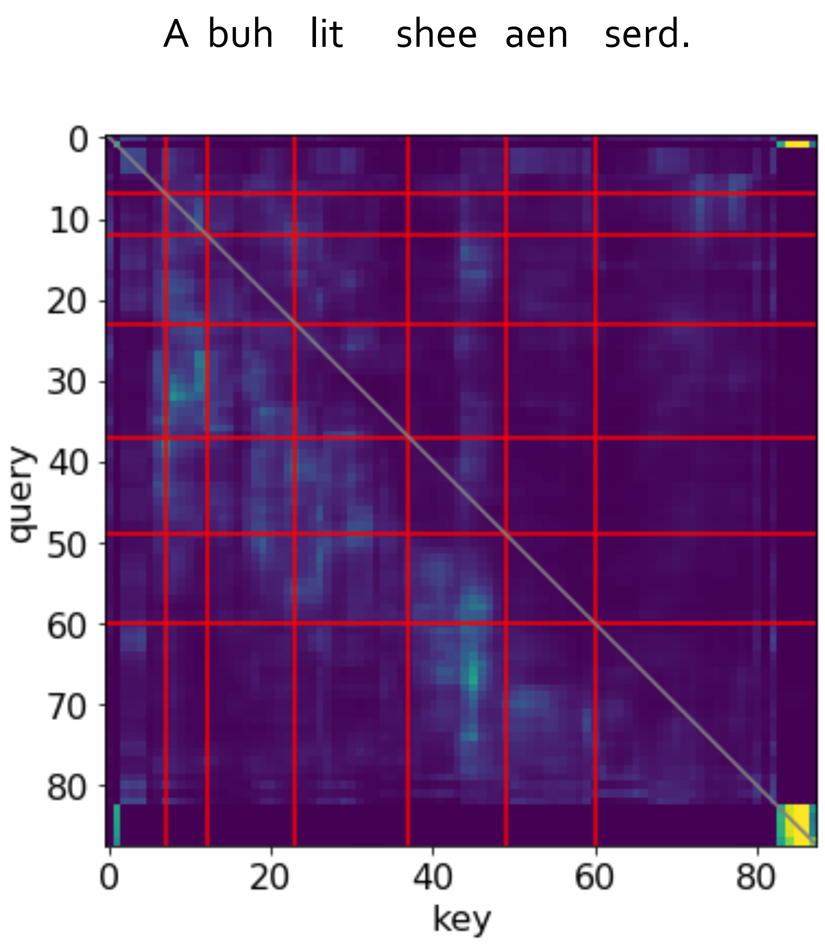
- Attention example: “A bullet, she answered. ”



Local attention



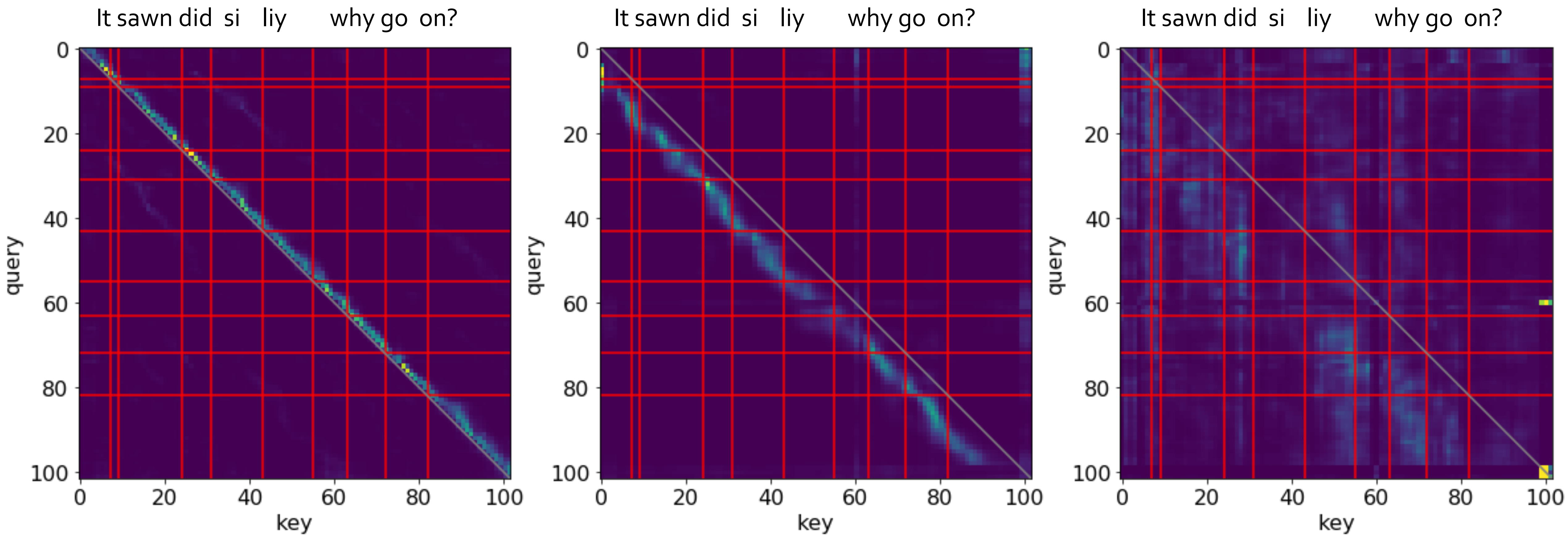
Attention to one syllable ahead



Attention to longer context

Context dependent computations

- Attention example: “It sounded silly, why go on?”

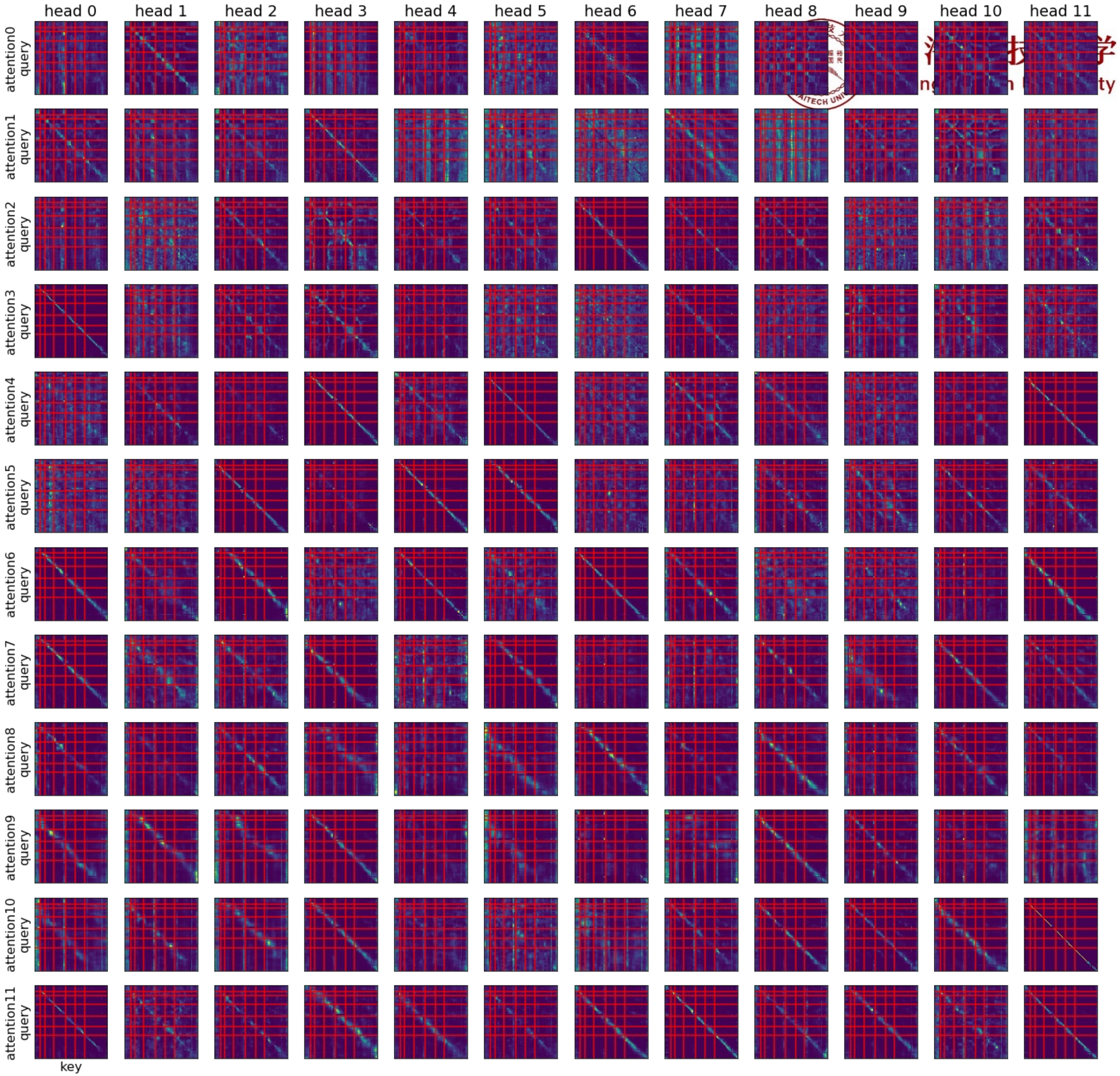


Local attention

Attention to one syllable ahead

Attention to longer context

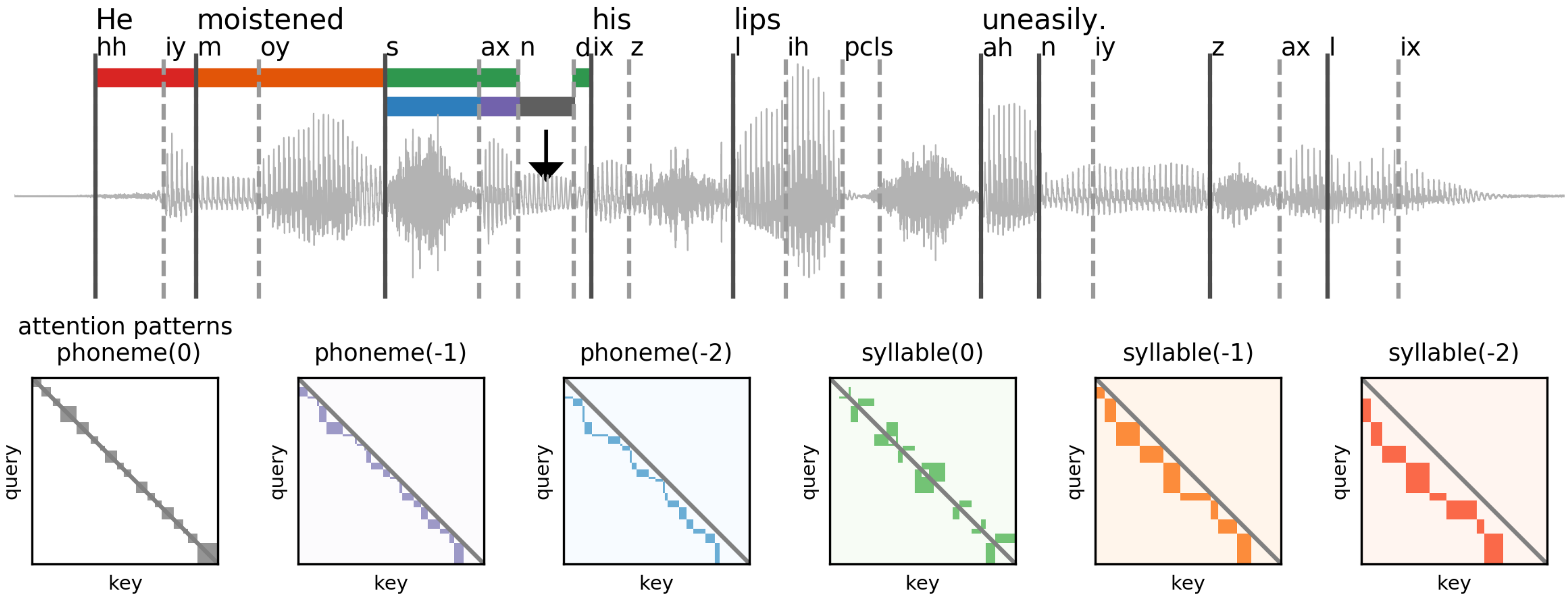
- Attention example
  - '"A bullet", she answered.'
  - HuBERT



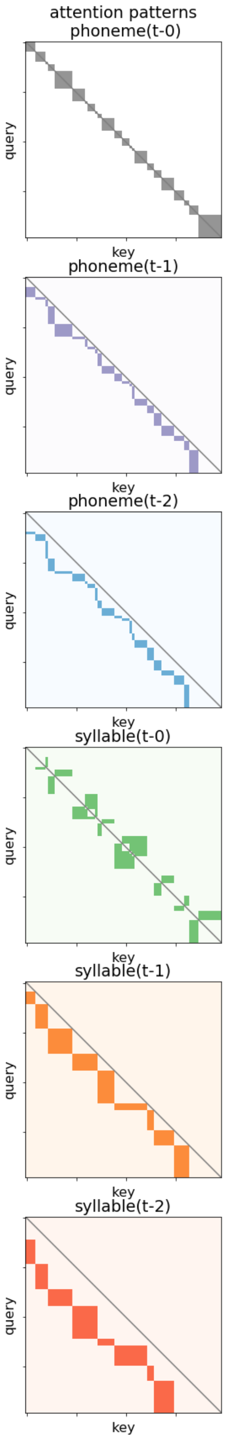
Parsing attentions according to temporal structures in speech

上海科技大学 ShanghaiTech University

- Attending to phonemic and syllabic context as stimulus-dependent computations



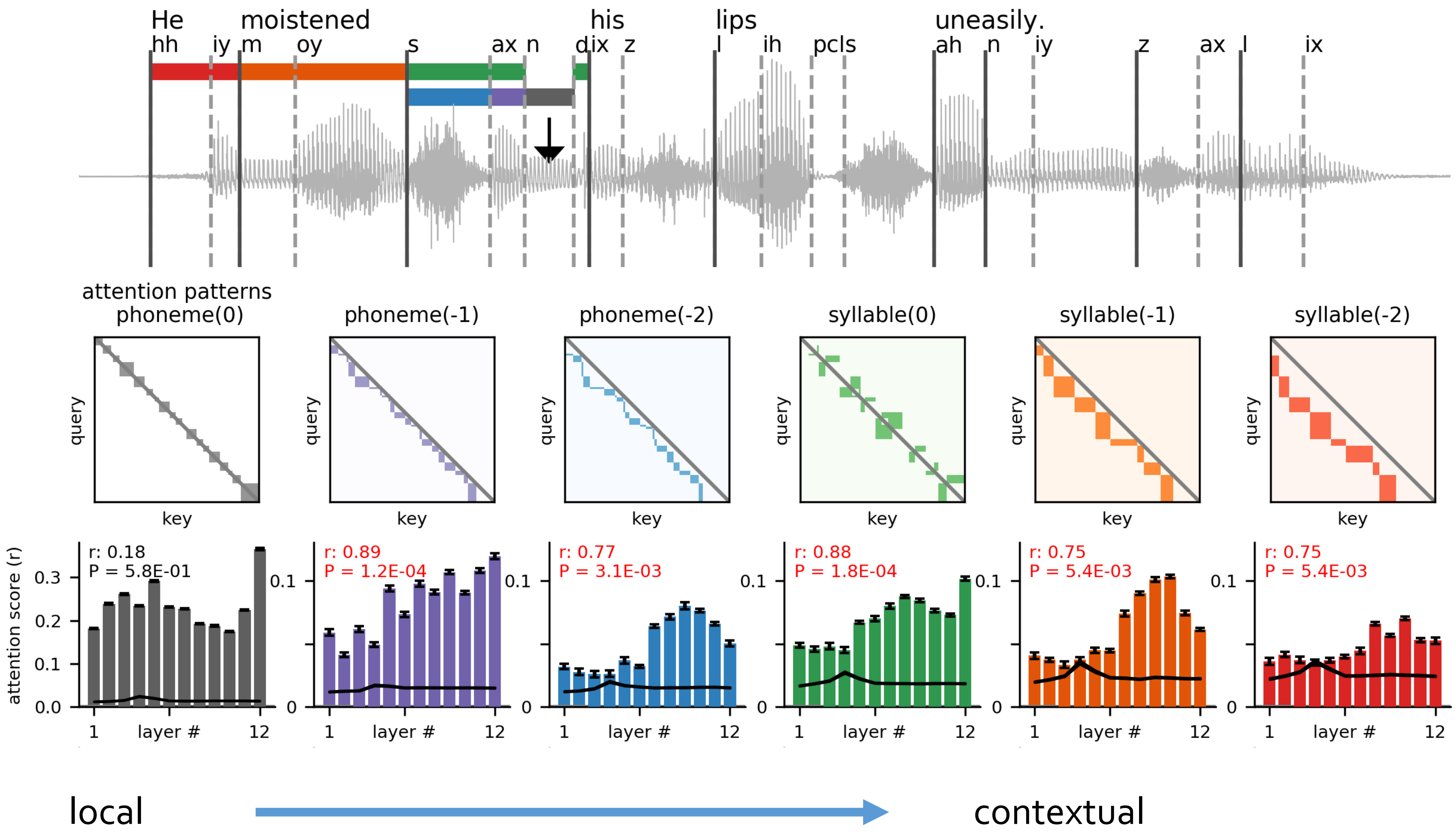
Increasing contextual dependency



Attention to phonemic and syllabic contexts

上海科技大学 ShanghaiTech University

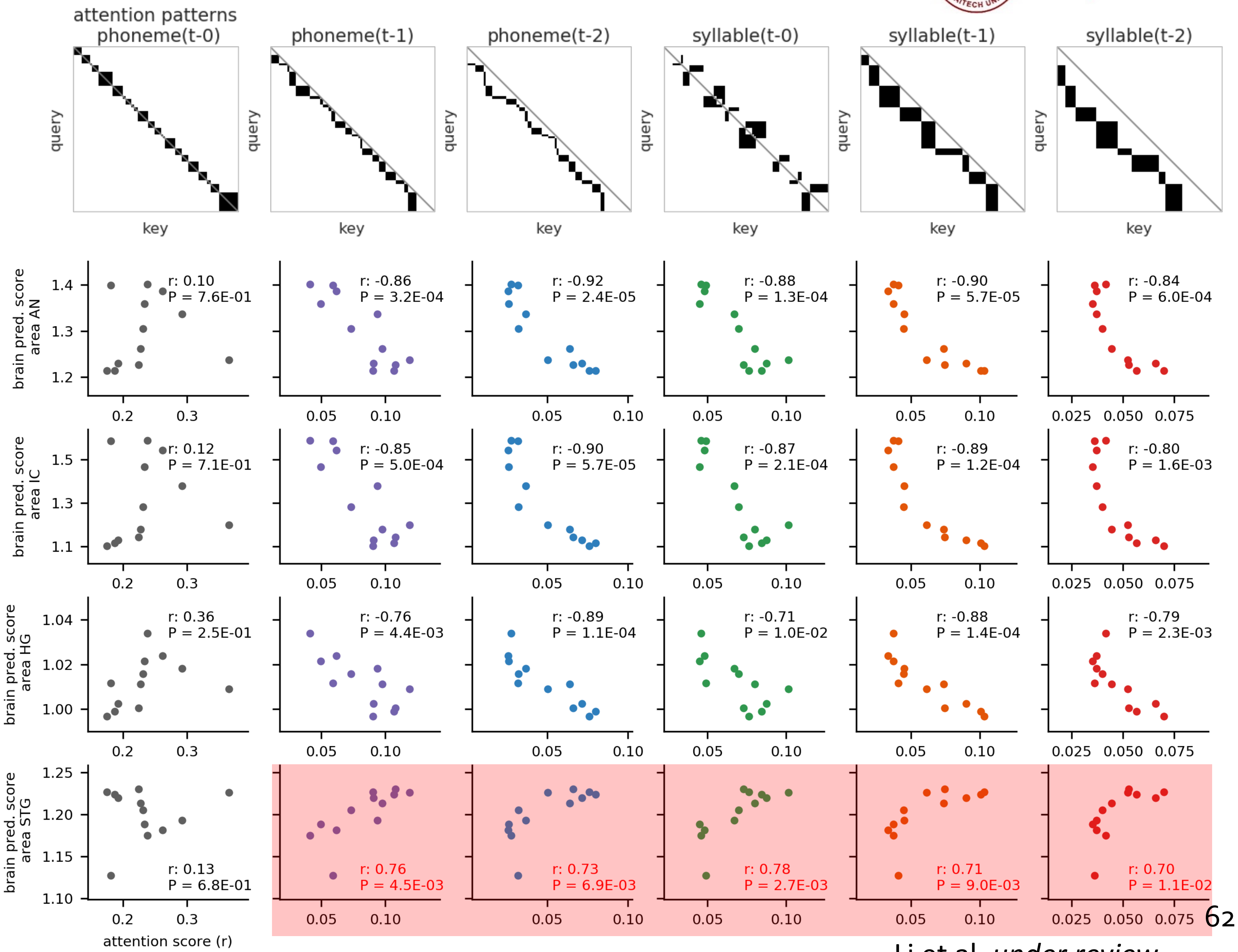
- Increased level of contextual phonemic and syllabic attentions along the hierarchy



Li et al. *under review*

Attention patterns explains brain correspondence

- Primary auditory cortex and auditory peripheral correspond to local phonemic computation
- STG corresponds to cross-phonemic and cross-syllable contextual attention

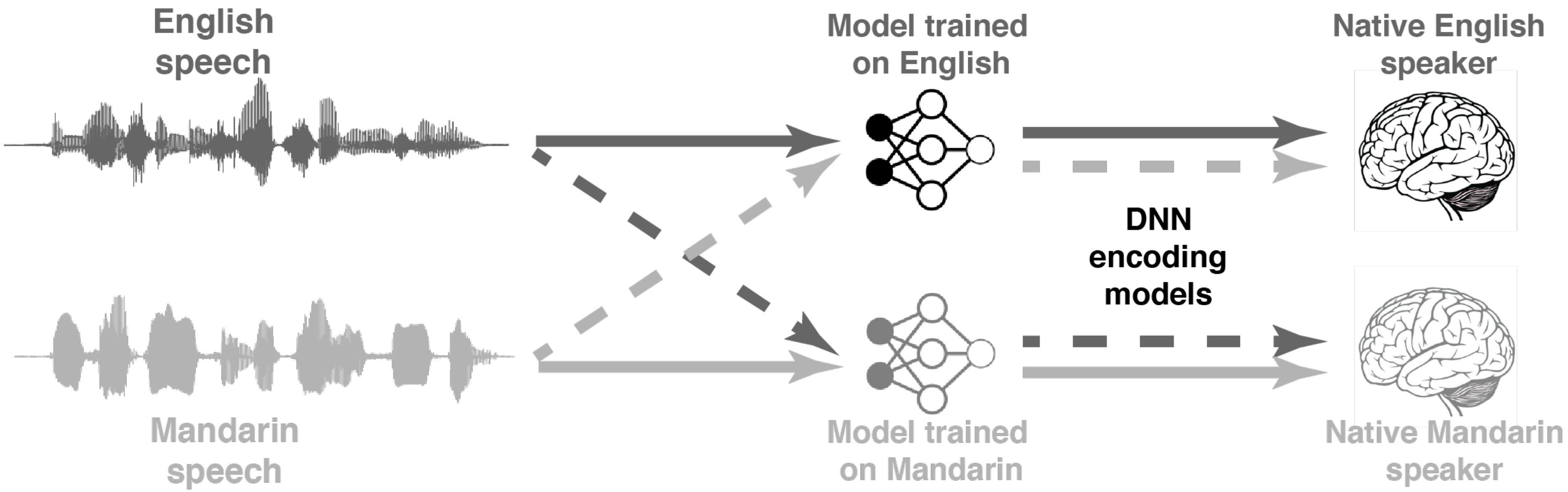


Li et al. *under review*

Language-specific representations & computations

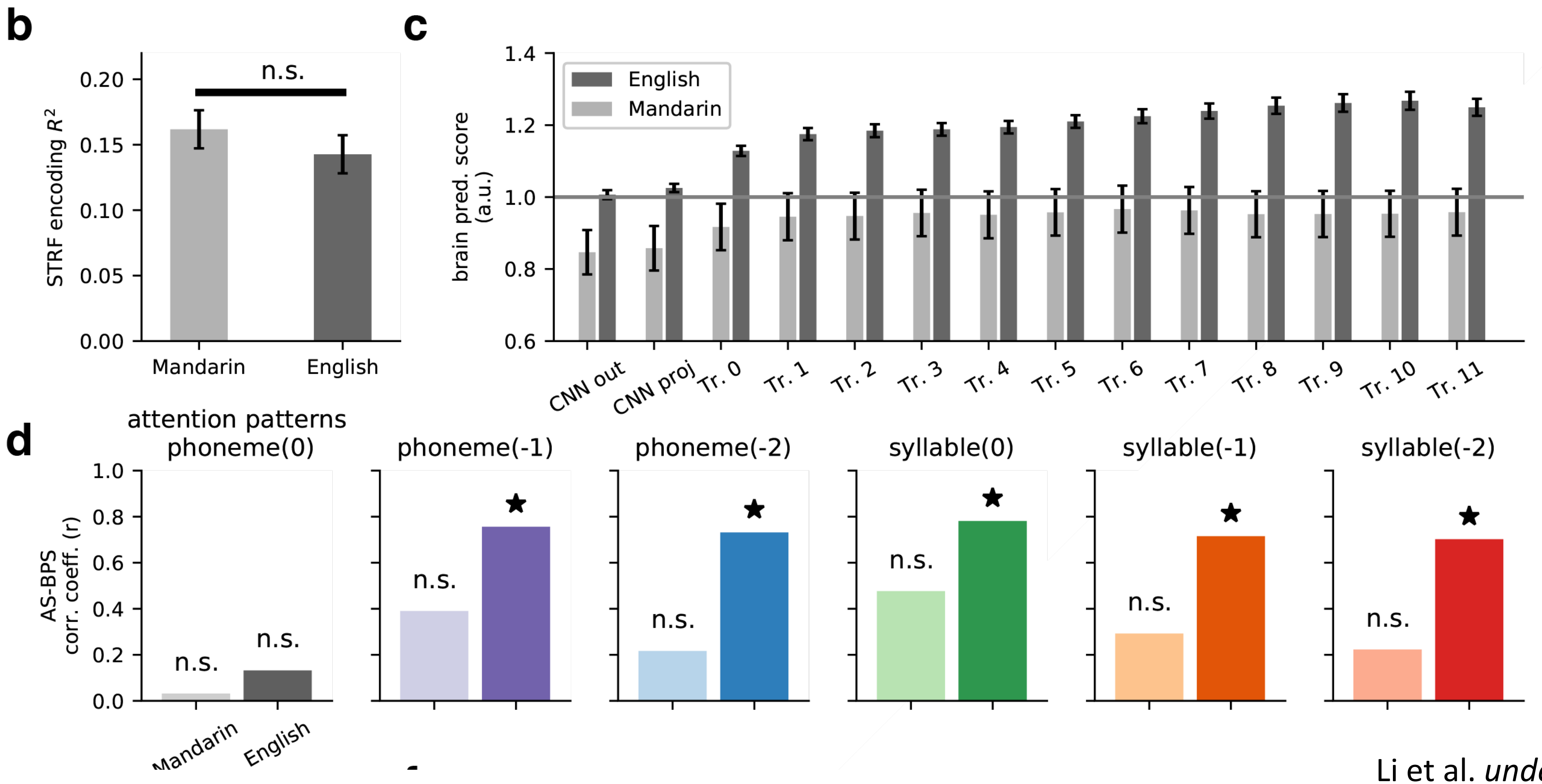
上海科技大学 ShanghaiTech University

- Cross-language comparisons in DNN and STG



Language-specific representations & computations

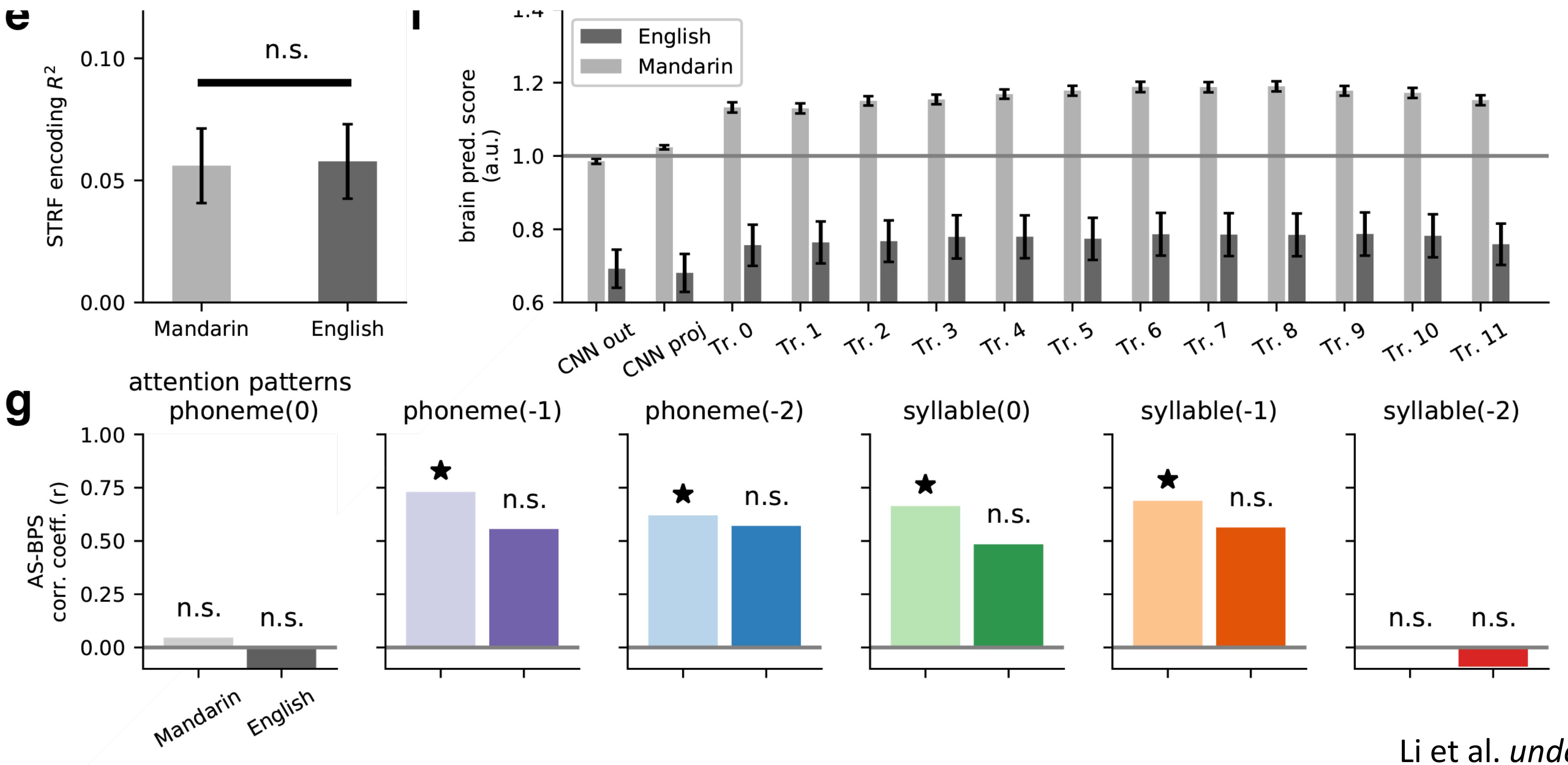
- STRF model is not sensitive to language-specific representations in STG of English speakers.
- English-pretrained model aligned to English speech better than Mandarin speech for native English speaker



Li et al. *under review*

Language-specific representations & computations

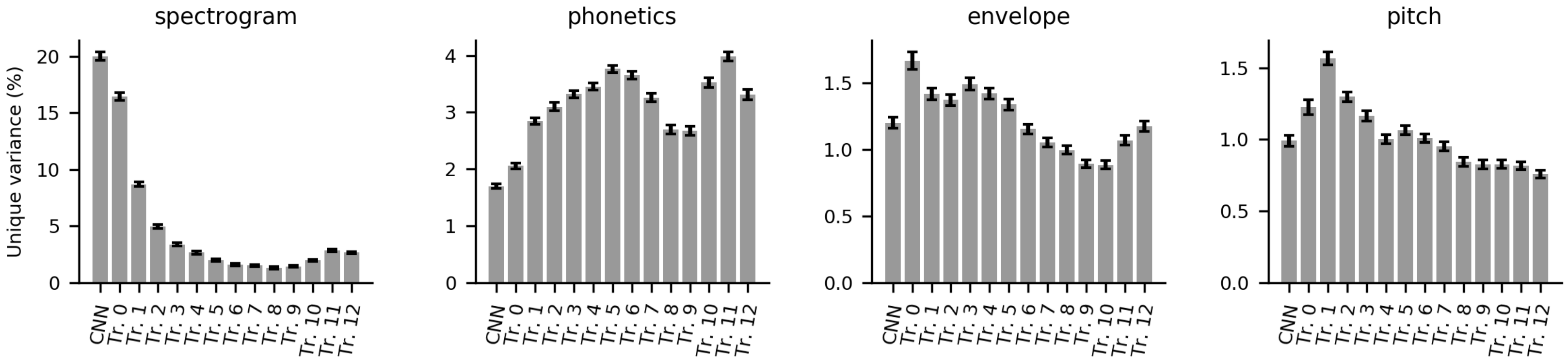
- Mandarin-pretrained model aligned to Mandarin speech for native Mandarin speaker

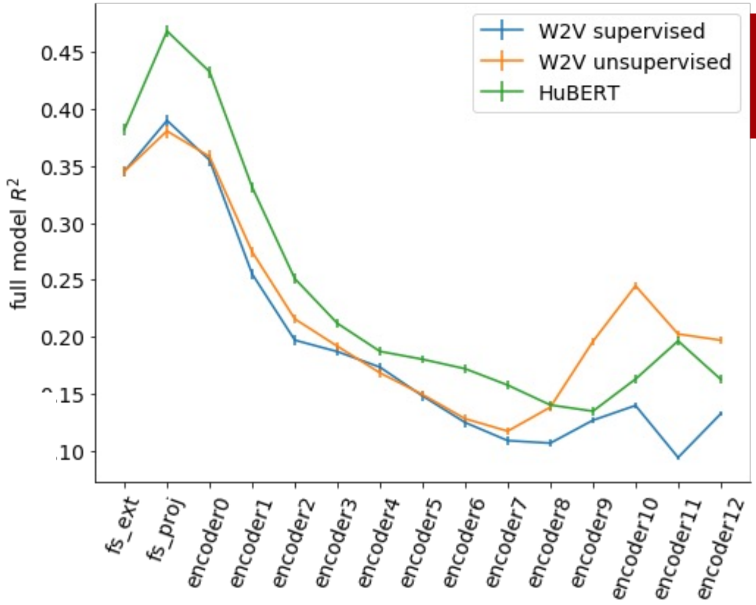


Li et al. *under review*

Feature representations

- Unique variance explained by each set of features in DNN
  - Spectro-phonological hierarchy

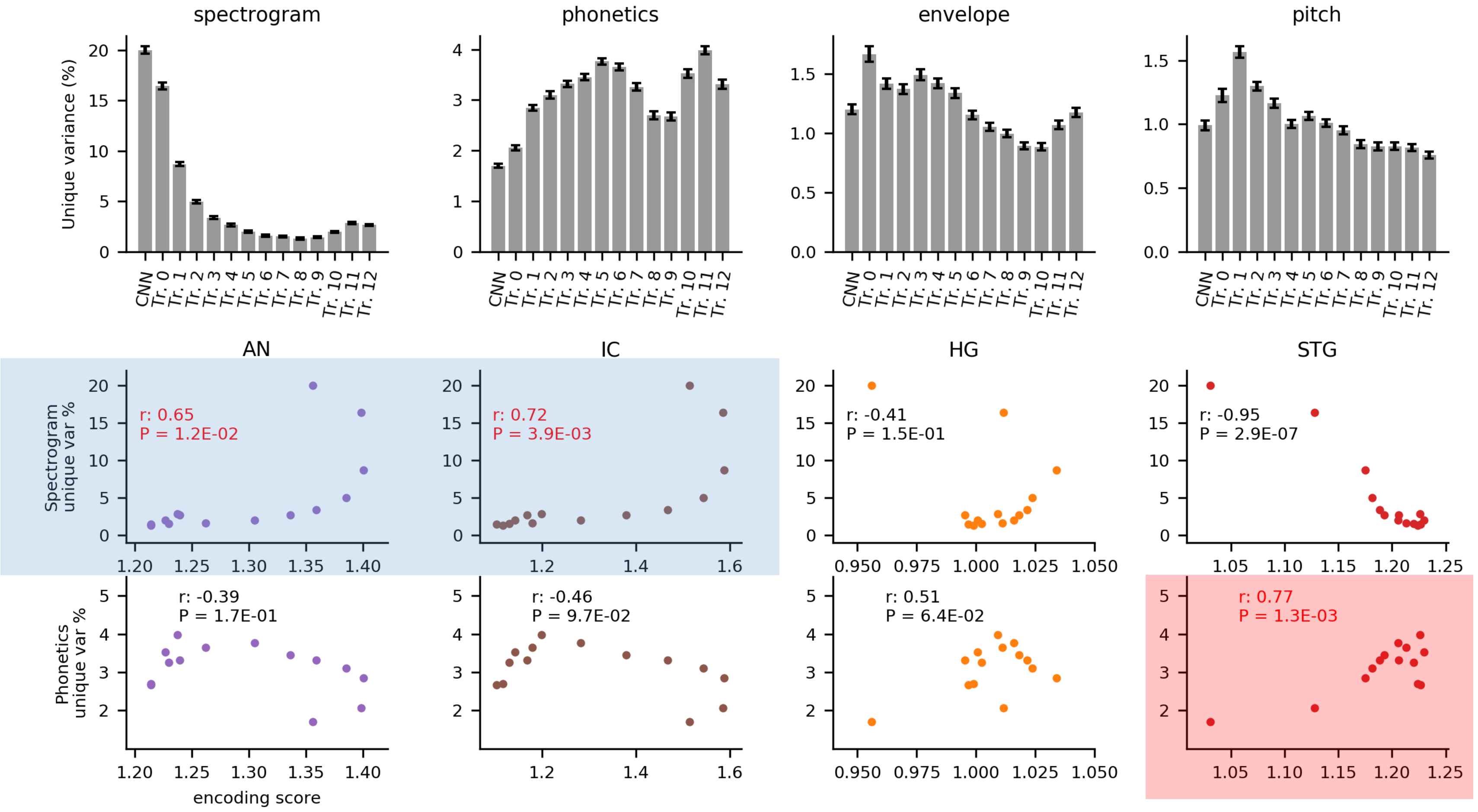


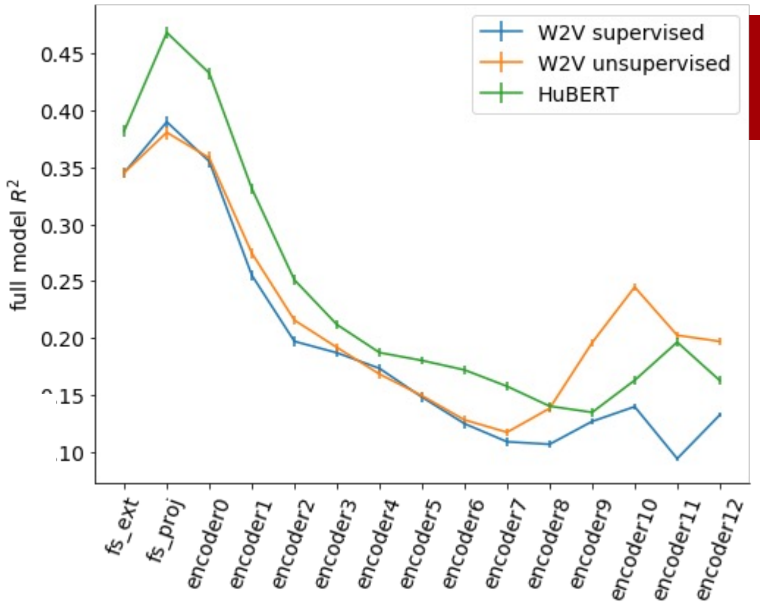


Li et al. *under review*

Feature representations

- Unique variance explained by each set of features in DNN
  - Spectro-phonological hierarchy





Li et al. *under review*

Research questions

- What is a good deep neural network model for speech perception in auditory pathway?

  - The early to later layers in the deep neural networks trained to learn speech representations correlate to the ascending AN-Midbrain-STG auditory pathway
  - Functional subpopulations in STG correlate to different contextual representation layers in DNN
  - The general results are consistent across network architecture and training objectives

- What are the key factors that make the DNN model good at predicting speech response in the brain?

  - Attention patterns explains brain correspondence: auditory pathway
  - Language-specific representation and computations aligned between DNN and STG
  - The representations in neural networks can be explained by an acoustic-phonological hierarchy

Open questions

- What is not captured by the DNN models and how to interpret it?
- How to incorporate top-down effects?
- Biological plausibility
- Higher-level information representation beyond phonetics

Marr’s three levels of analysis

David Marr’s three levels

- ■ Computational level
  - ➢ cognitive task and problem
- ■ Algorithm level
  - ➢ information representation and transformation
- ■ Implementation level
  - ➢ implementation through interactions between basic elements

Neural coding

Computational model

- Fit the underlying neural activity
- Perform the cognitive tasks

AI models can do both!

Marr 1982

Marr’s three levels of analysis

David Marr’s three levels

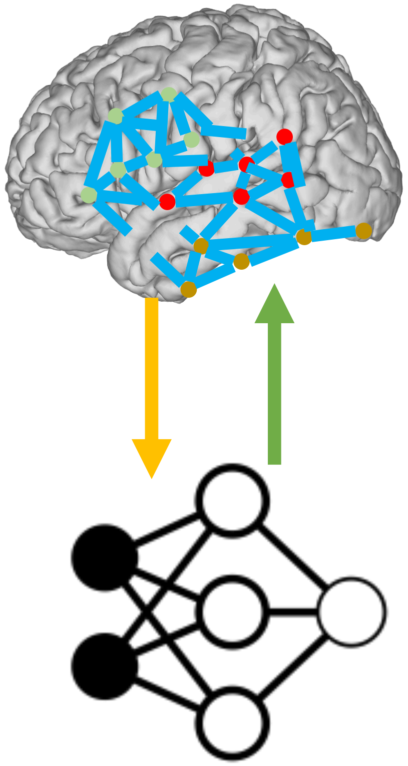
- ■ Computational level
  - ➢ cognitive task and problem
- ■ Algorithm level
  - ➢ information representation and transformation
- ■ Implementation level
  - ➢ implementation through interactions between basic elements

Neural coding

Computational model

Brain-inspired AI model

Brain network



AI models

Marr 1982

Thank you!

yuanningli@gmail.com
https://yuanningli.github.io/

Demo Code

上海科技大学 ShanghaiTech University

- GitHub: https://github.com/yuanningli/neural_encoding_demo
- QR code:

yuanningli@gmail.com
https://yuanningli.github.io/