CAI WINGFIELD

DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF CAMBRIDGE

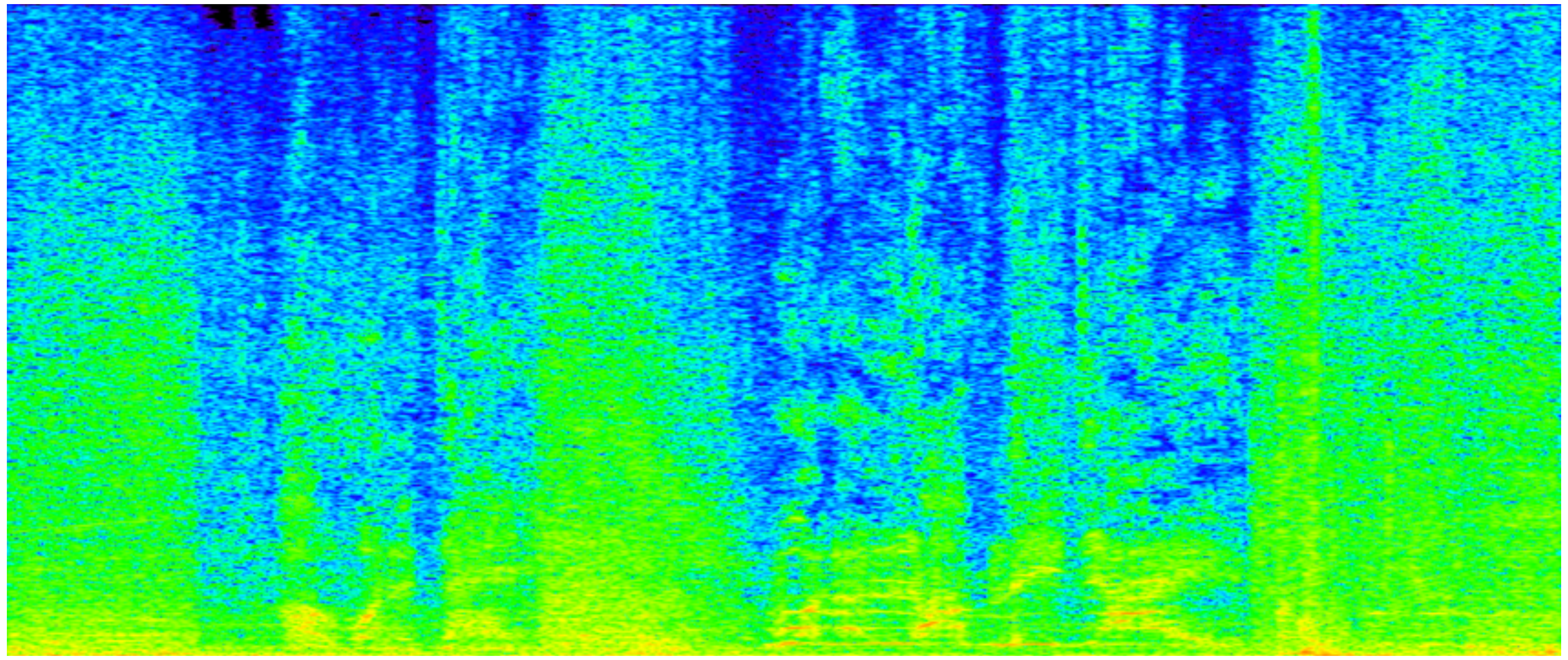
CARDIFF METROPOLITAN UNIVERSITY
17 NOVEMBER 2015

AUTOMATIC SPEECH RECOGNISER REVEALS PHONETIC FEATURE REPRESENTATIONS IN HUMAN AUDITORY CORTEX

THE QUESTION

HOW DOES THE HUMAN BRAIN EXTRACT MEANING FROM HEARD SPEECH?

SPOKEN LANGUAGE COMPREHENSION IN HUMANS



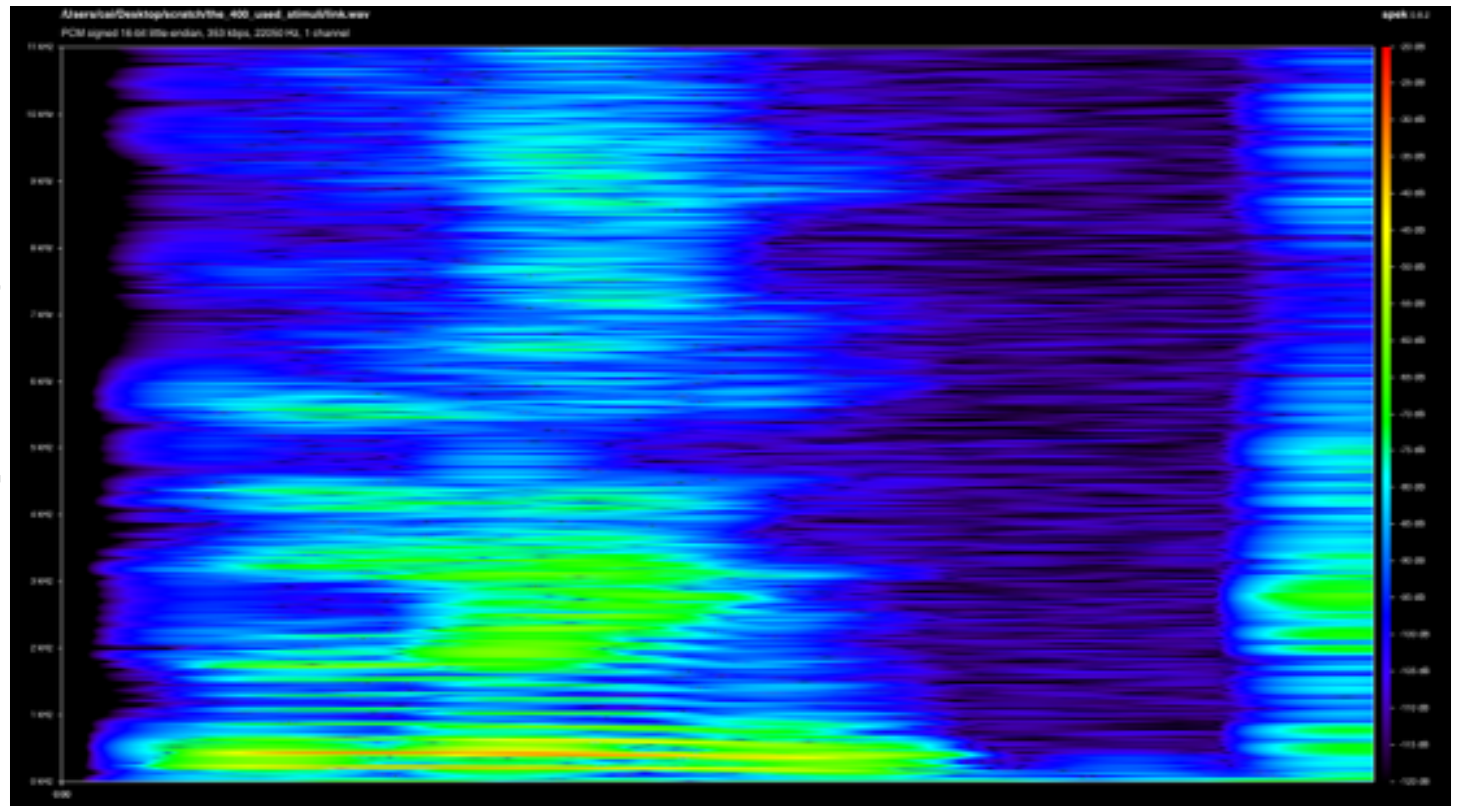
Noisy, continuous speech input

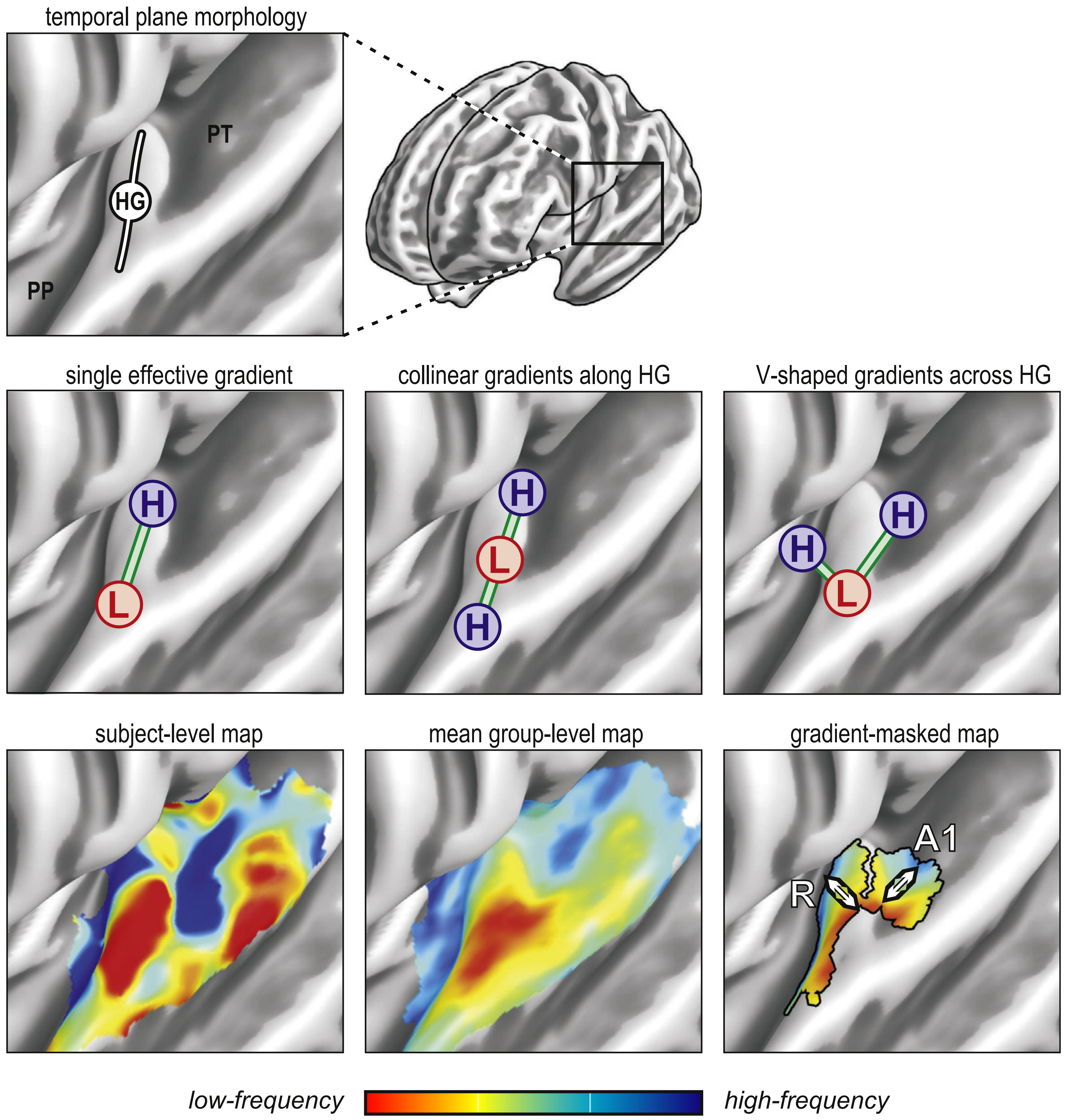
"what a lovely day"

Abstract word identities

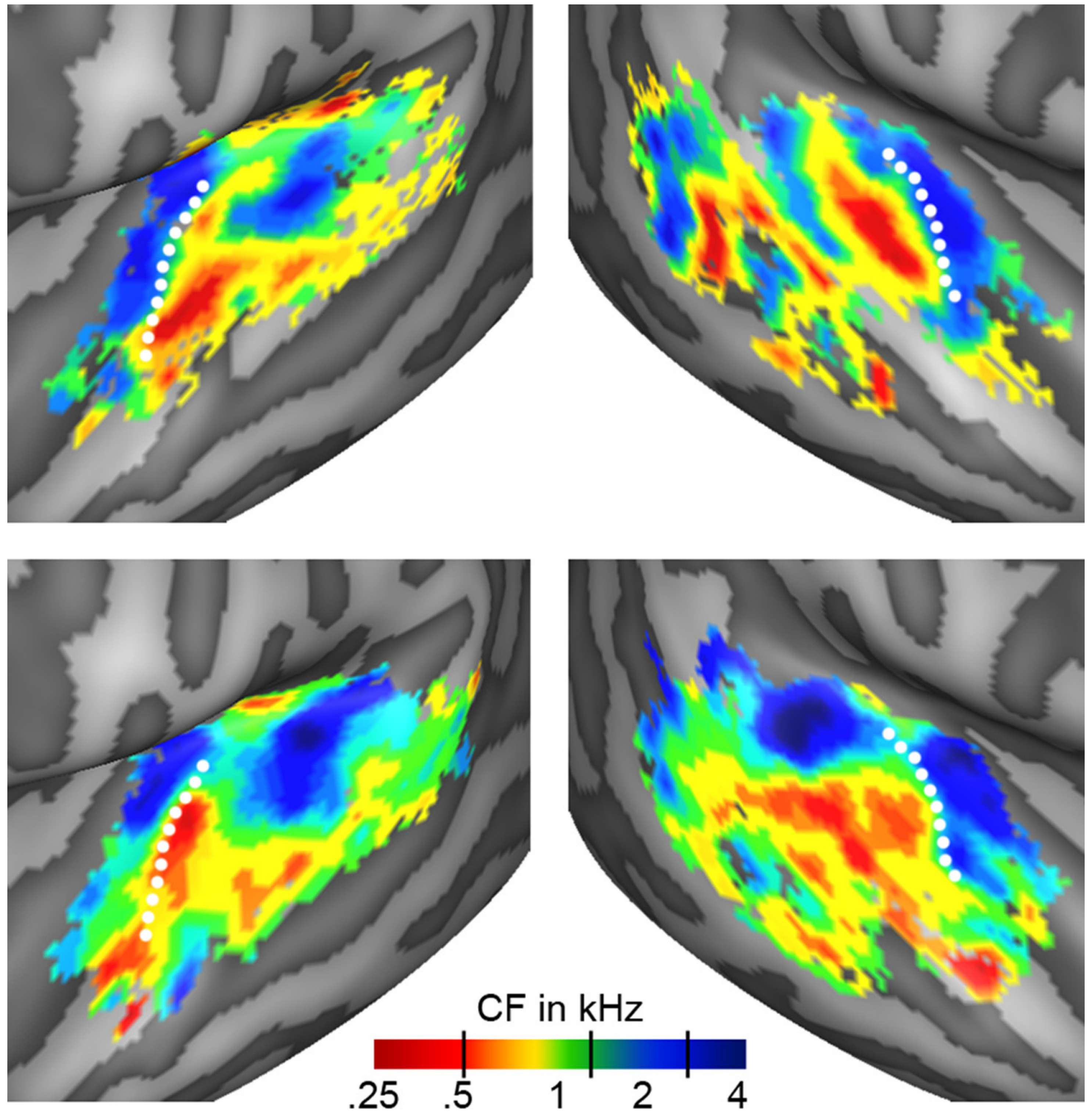
FREQUENCY-RELATED INFORMATION IN THE BRAIN

"L I N K"





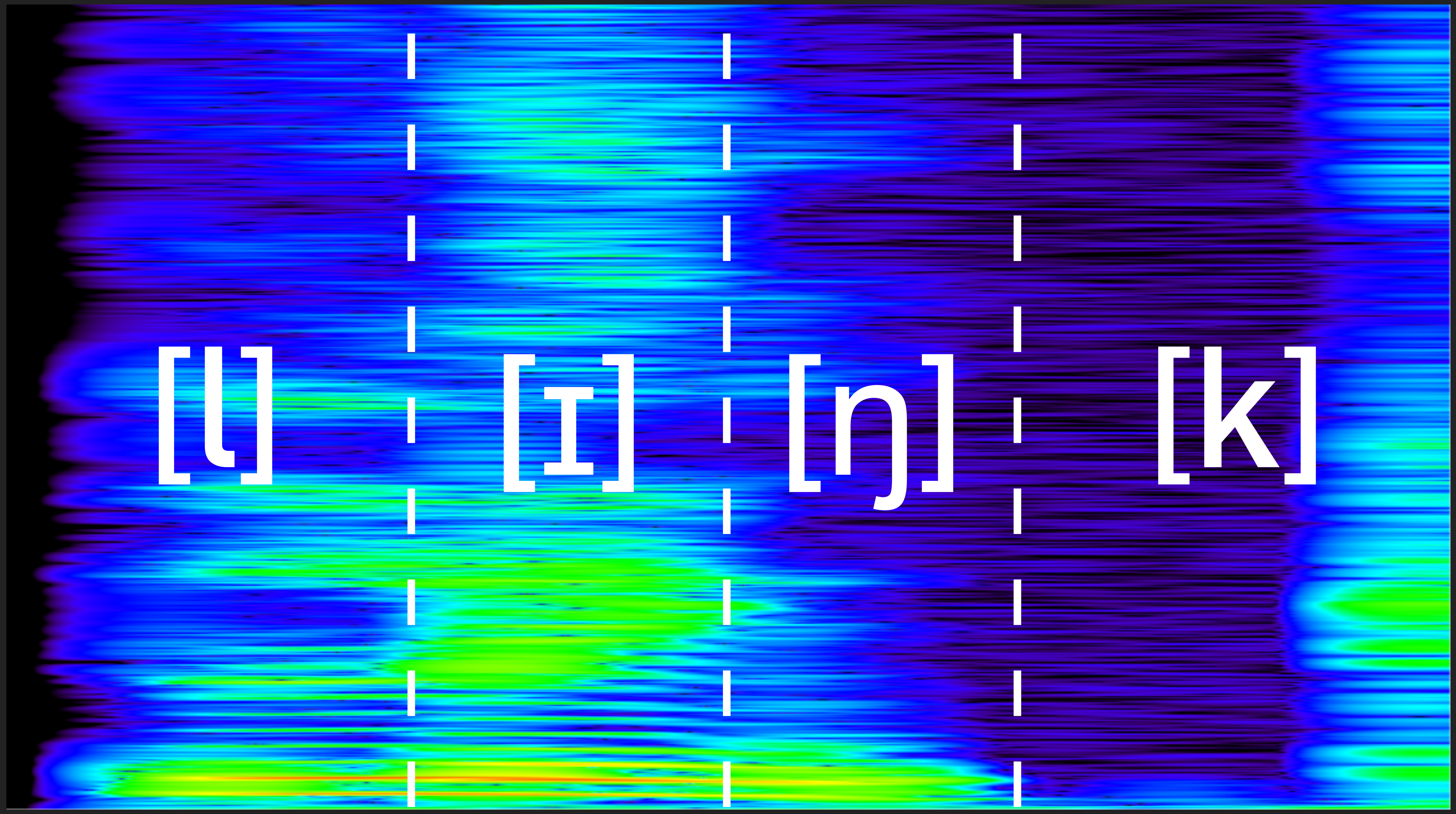
Saenz & Langers (2014)
Hearing Research



Moerel et al. (2012)
Journal of Neuroscience

THE BRAIN EXTRACTS MEANING FROM SOUND

- The brain receives raw acoustic input from the ears.
- The brain perceives individual words in continuous speech.
- Some complex neurobiological processes analyse features of the speech to extract meaning.



“L I N K”

AUTOMATIC SPEECH RECOGNITION (ASR)

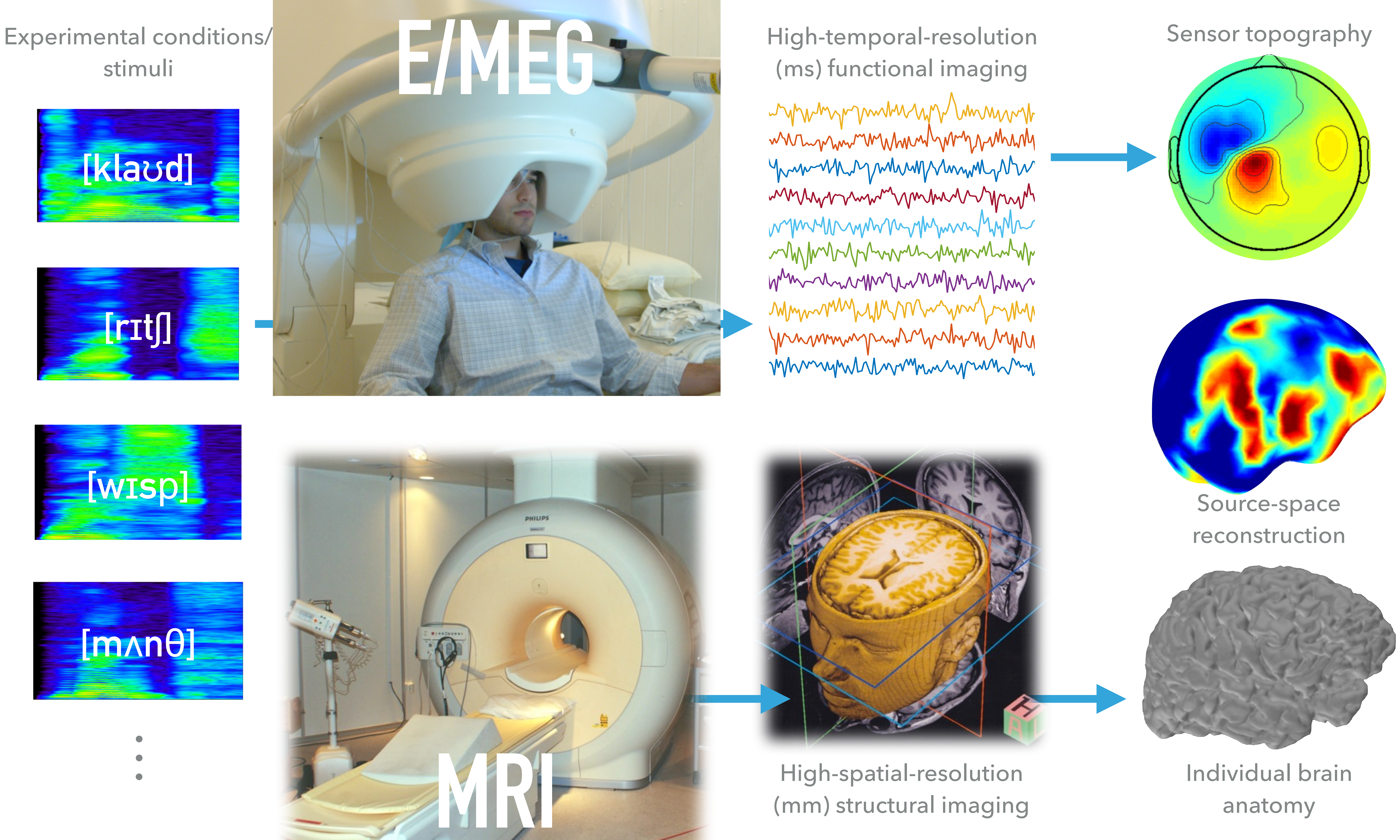
- Software-based ASR systems perform the same task as humans.
  - Speech goes in, words come out.
- They provide a computation account of how the task can be achieved.
- We will use their intermediate-level representations to model feature processing in the brain.

What kind of features would we expect to find?

How can we compare machine states to brain states?

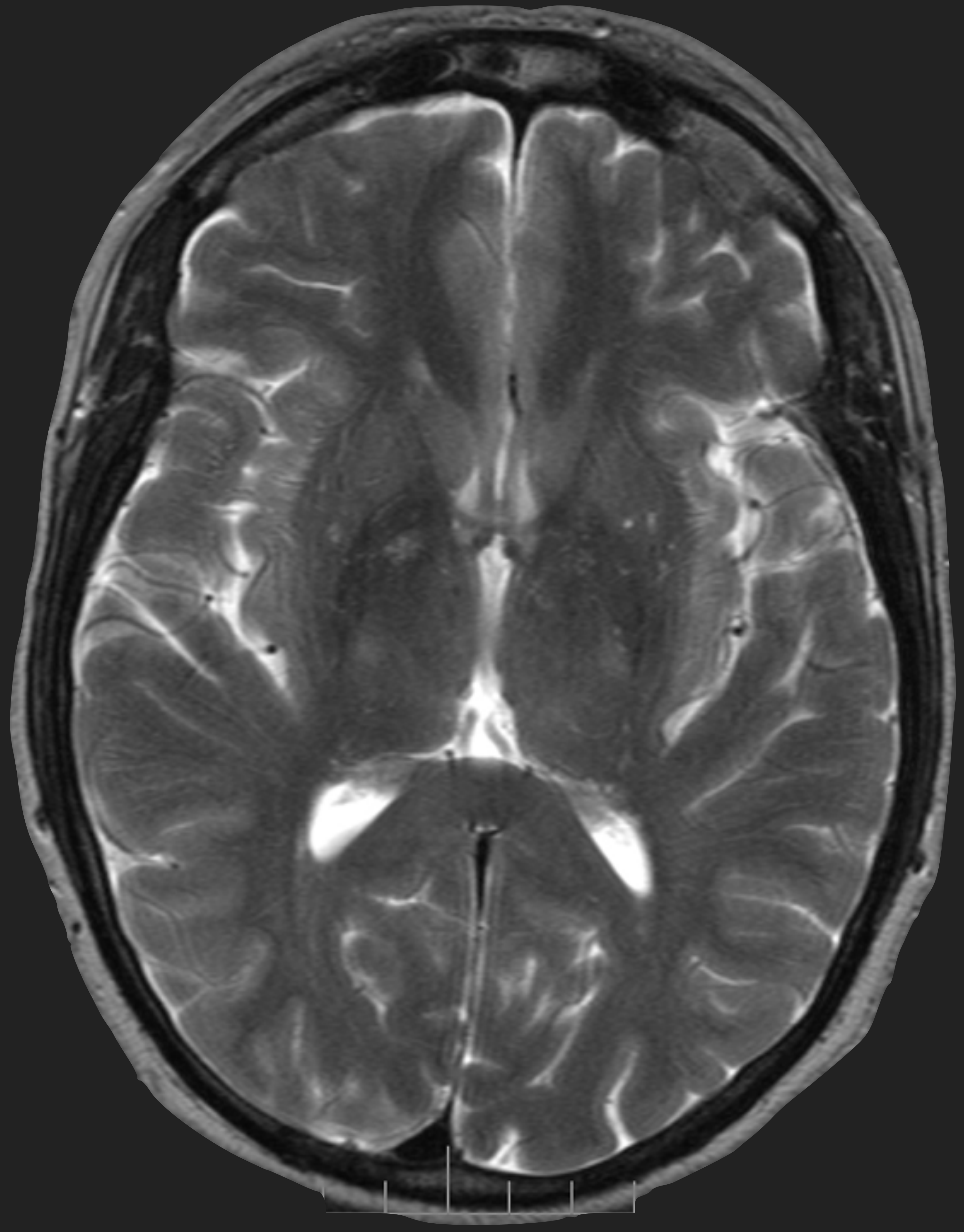
FUNCTIONAL NEUROIMAGING

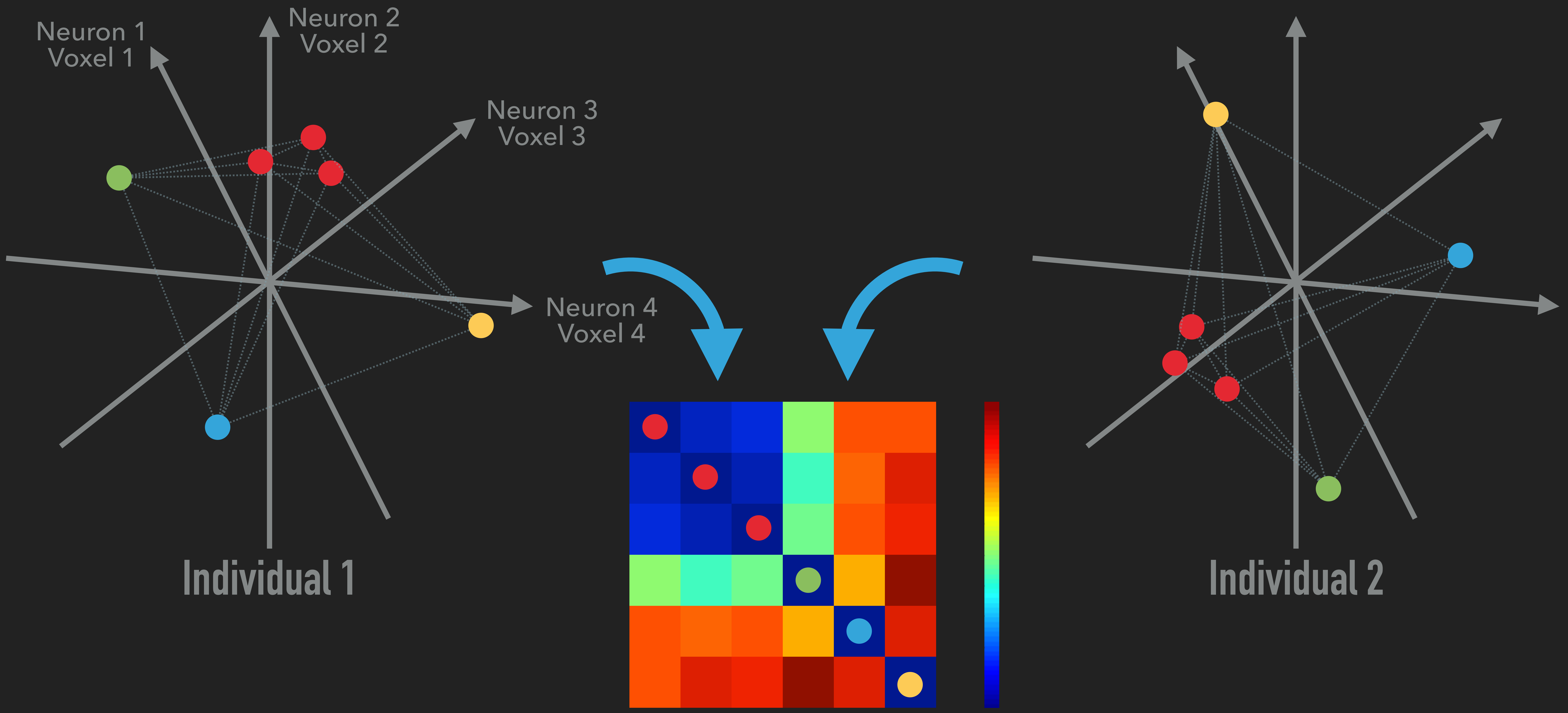
INVESTIGATING HOW AND WHERE THE BRAIN REPRESENTS INFORMATION


Experimental conditions/
stimuli
[klaʊd]
[rɪtʃ]
[wɪsp]
[mʌnθ]
E/MEG
MRI
PHILIPS
High-temporal-resolution
(ms) functional imaging
High-spatial-resolution
(mm) structural imaging
Sensor topography
Source-space
reconstruction
Individual brain
anatomy

COMPARING REPRESENTATIONS

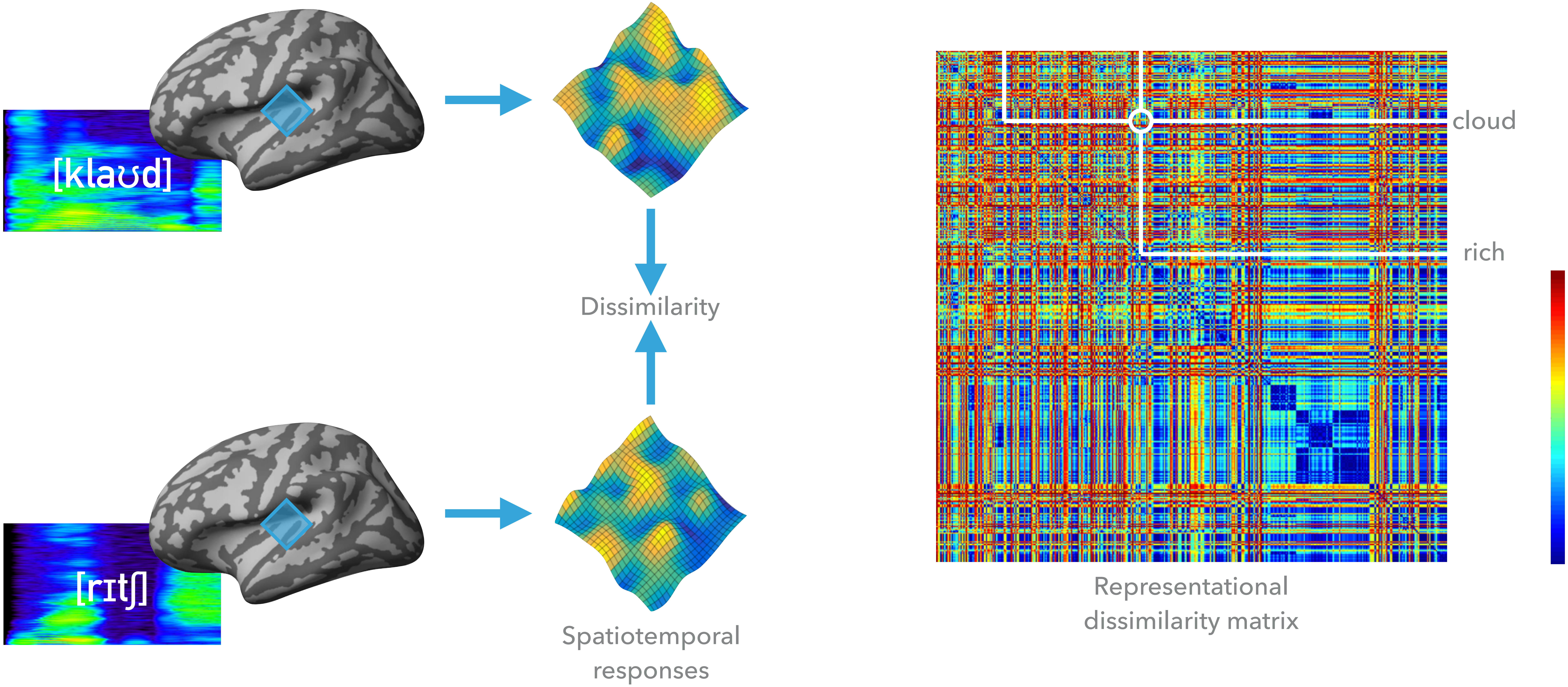
- Can't assume fine-scale correspondence between subjects.
- Can't assume any information present will be of the same format.
- Instead, we look at individual representations: Reproducible patterns in localised activity.





Kriegeskorte:
“Representational geometries”

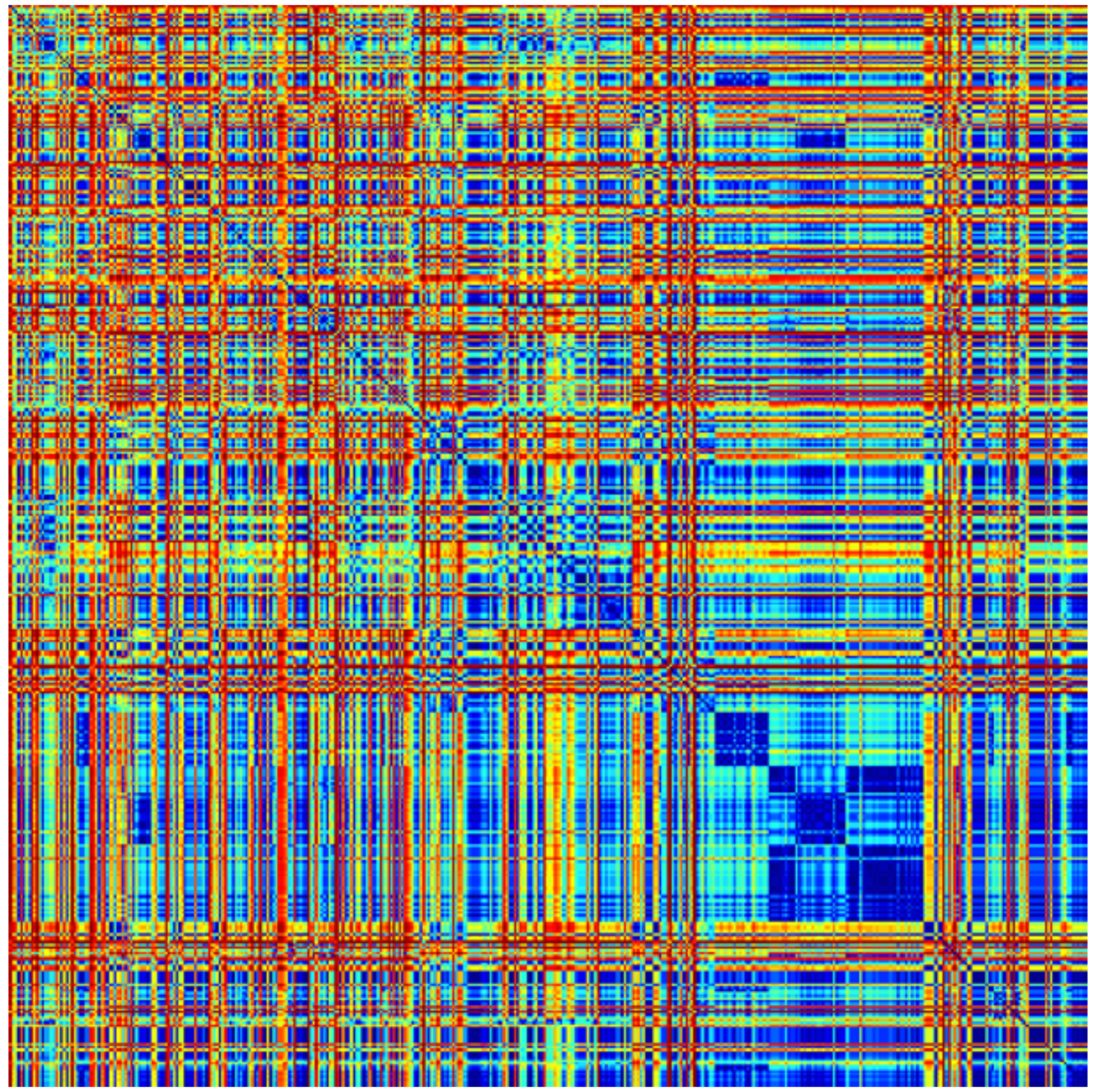
REPRESENTATIONAL SIMILARITY ANALYSIS

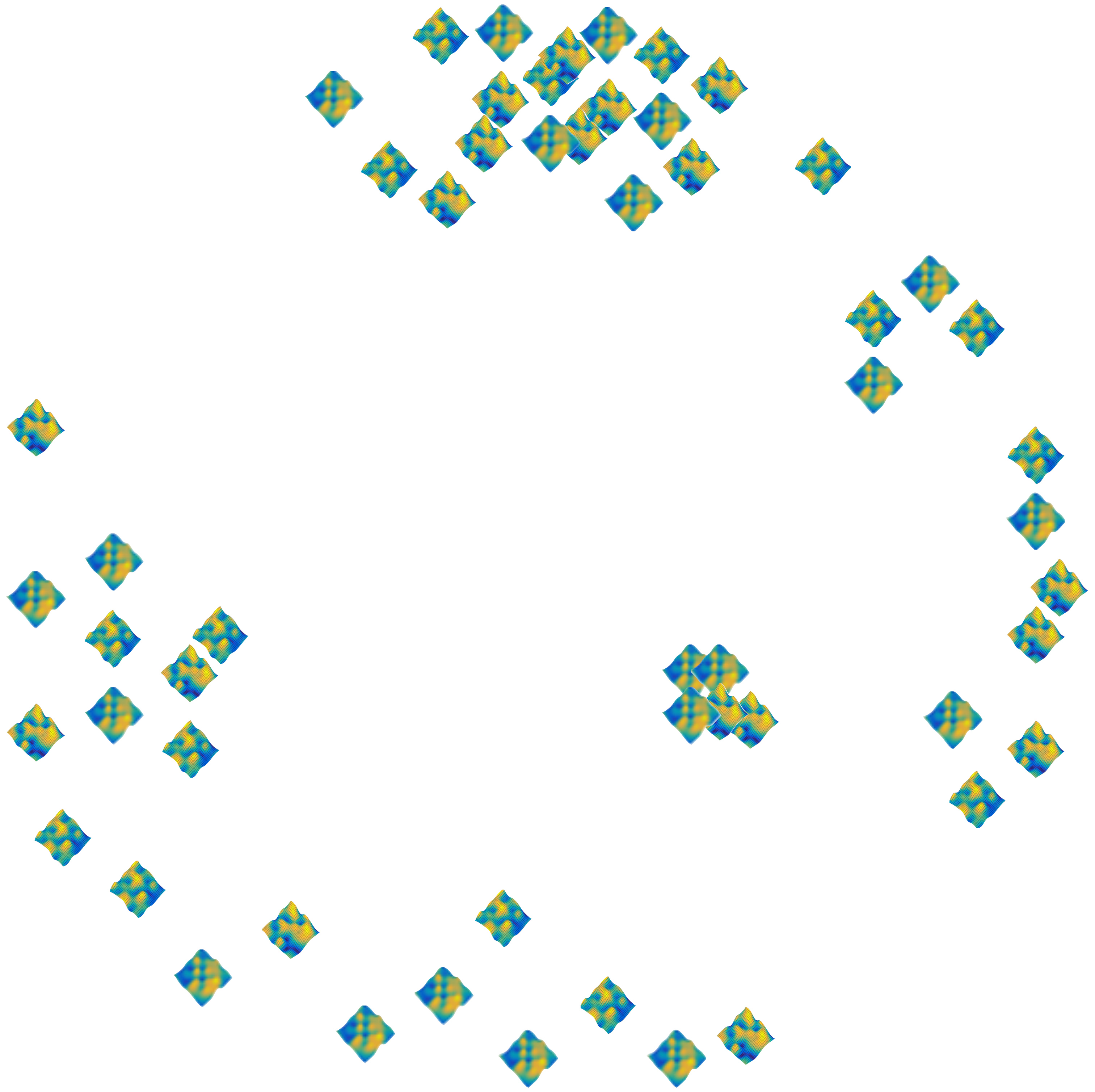


REPRESENTATIONAL SIMILARITY ANALYSIS

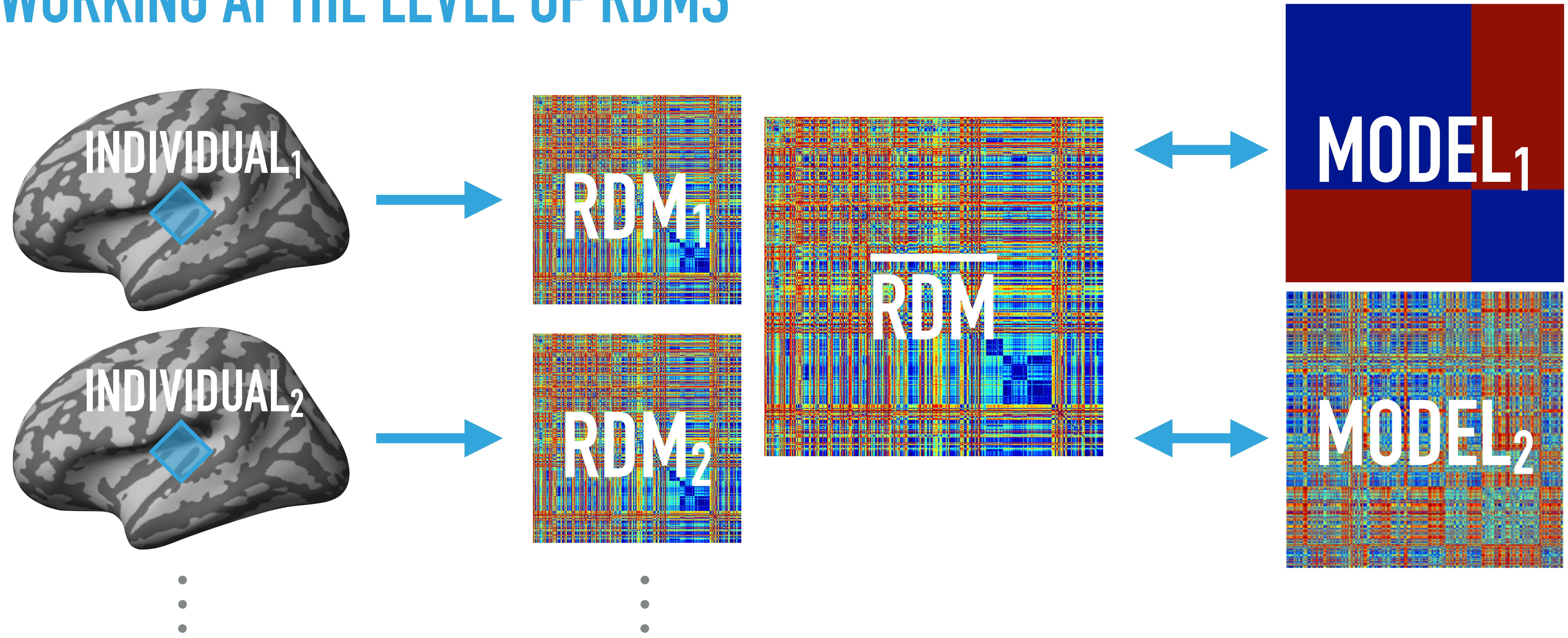
Kriegeskorte et al. (2008)
Frontiers in Systems Neuroscience

- Dissimilarities between responses characterise a representational space.
- Treated as a distance matrix, we can see how a brain region "views" the stimulus set.





WORKING AT THE LEVEL OF RDMS



Category model

Computational model

Unable to compare individuals' responses

Combining loses fine-grained information

Able to compare individuals' RDMs

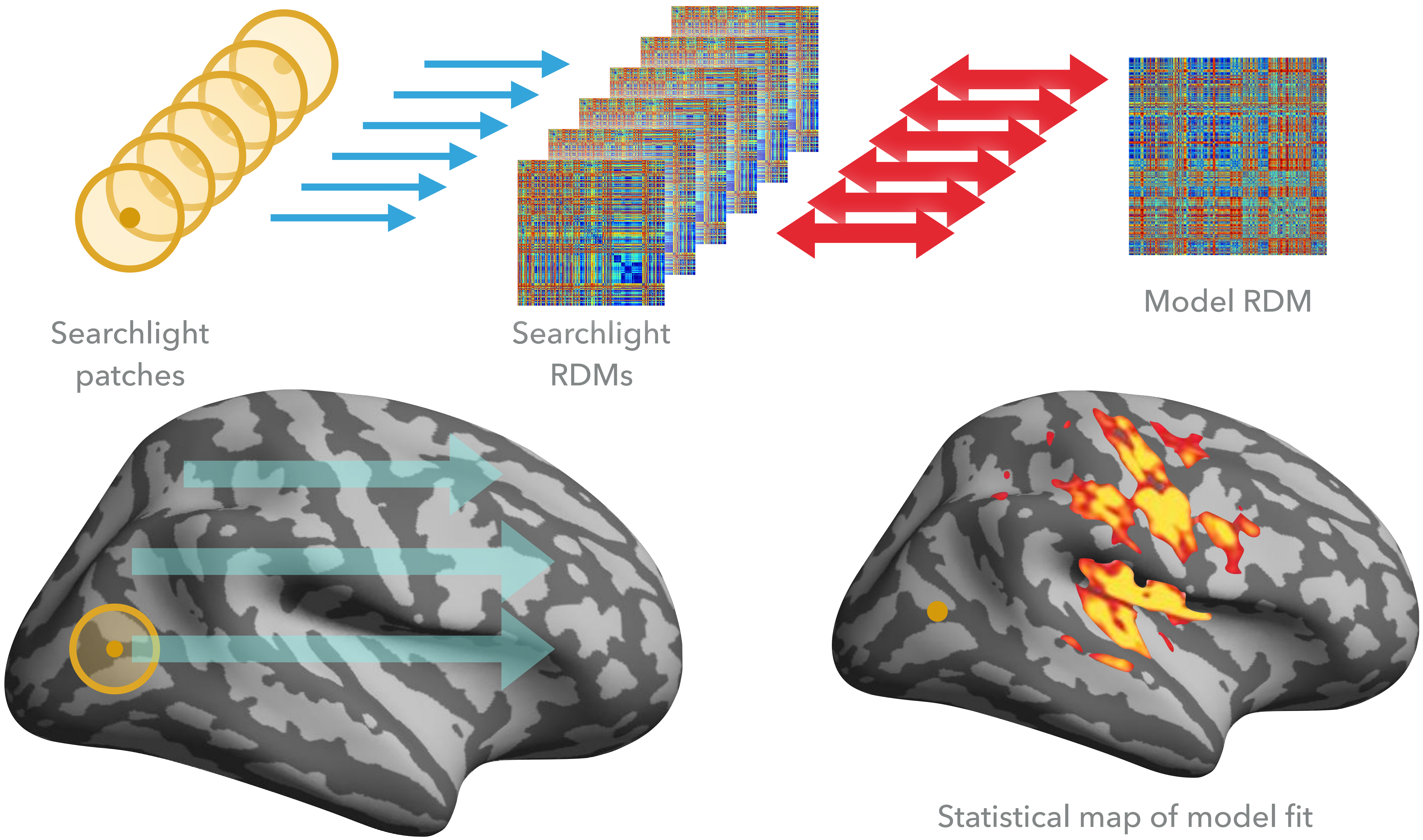
Combining preserves fine-grained information

Can test hypotheses about representational space

Kriegeskorte et al. (2006)
PNAS

Su et al. (2012)
International Workshop on Pattern Recognition in NeuroImaging

SEARCHING FOR MODEL FIT: SEARCHLIGHT RSA



- Take brain data from a regular "searchlight".
- Compute 1 RDM from all data inside that region.
- Match each RDM to a fixed model.
- Statistical brain map of information.

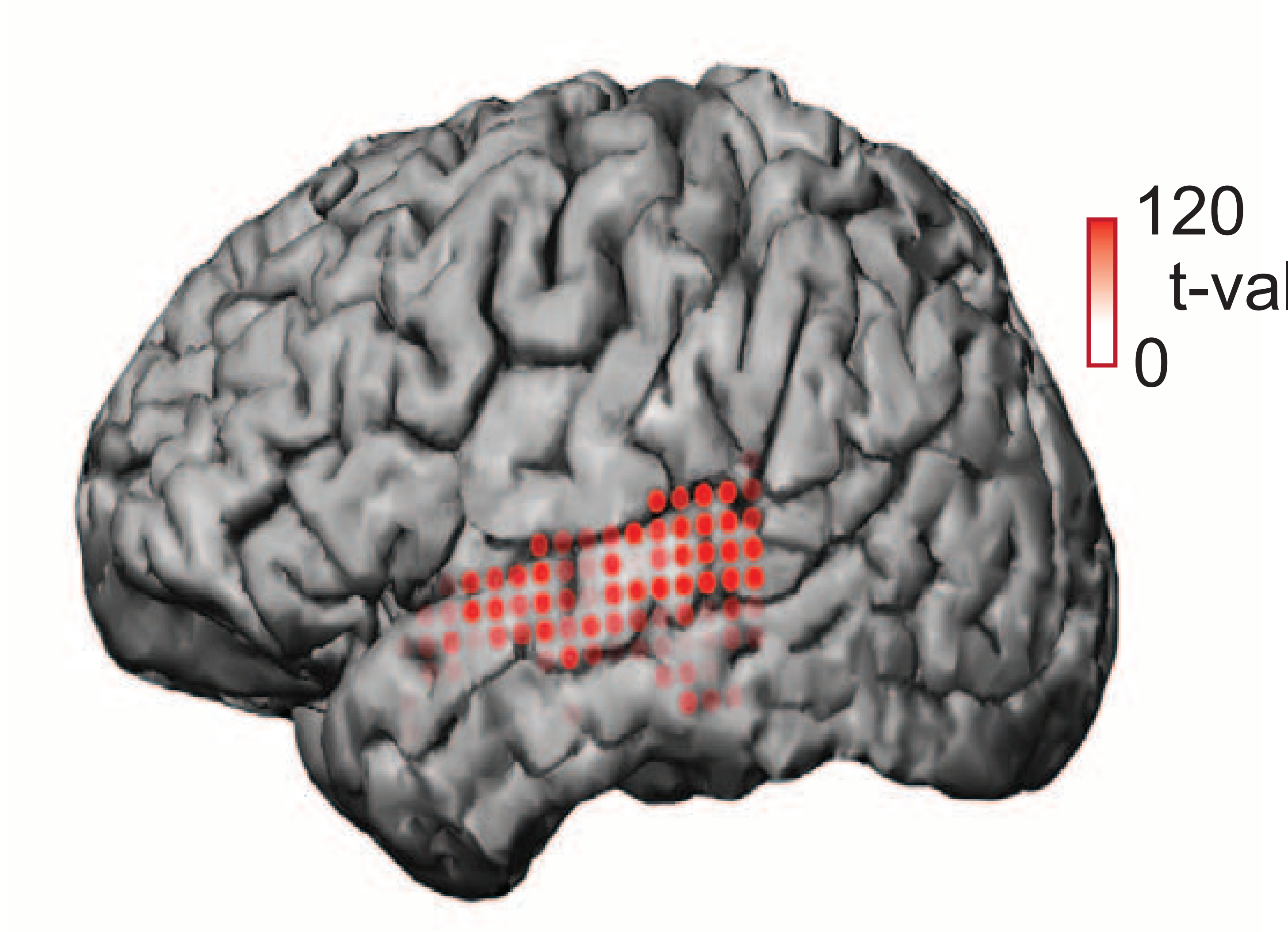
MODELLING SPEECH RESPONSES:

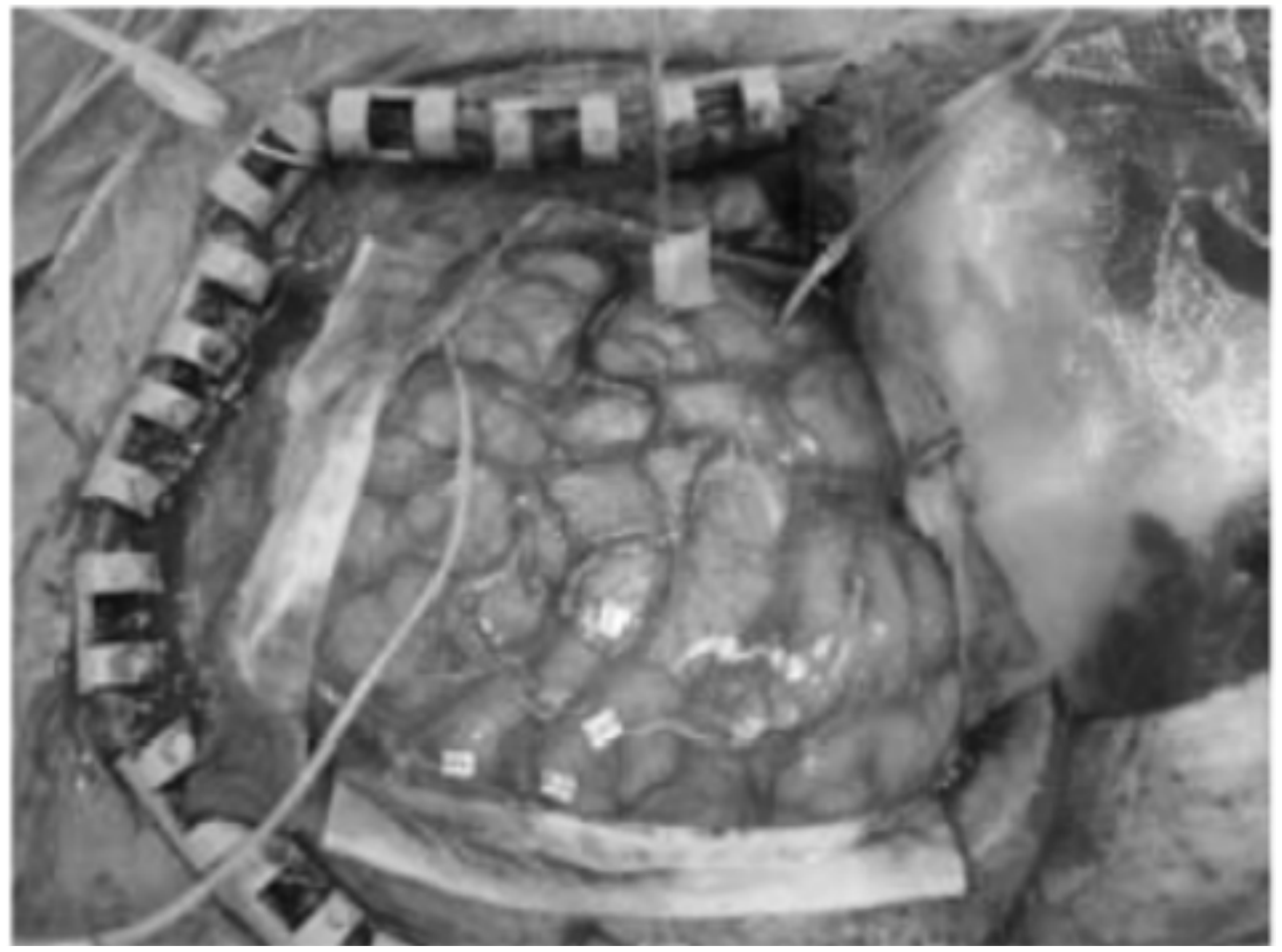
PHONES AND PHONETIC FEATURES FROM AN ASR SYSTEM

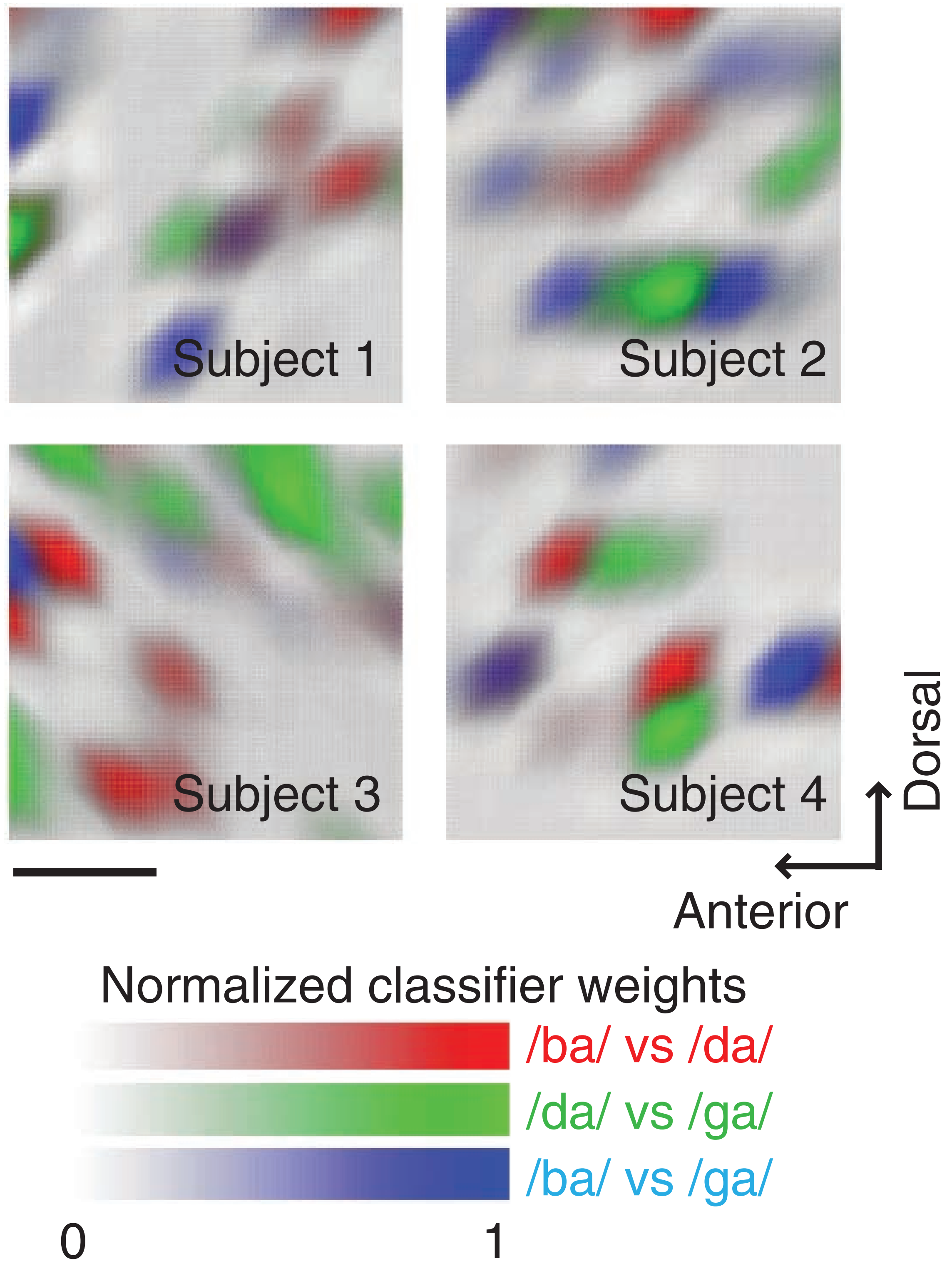
PHONEMES, PHONES AND FEATURES

- **Phonemes** are parts of speech which distinguish words in a language.
    - /l/ and /r/ in English, not in Japanese.
- **Phones** are parts of speech produced in a distinct manner.
    - No English words differ only by [r] vs [ɹ].
- **Articulatory features** are ways of classifying phones based on the place and manner of their articulation.

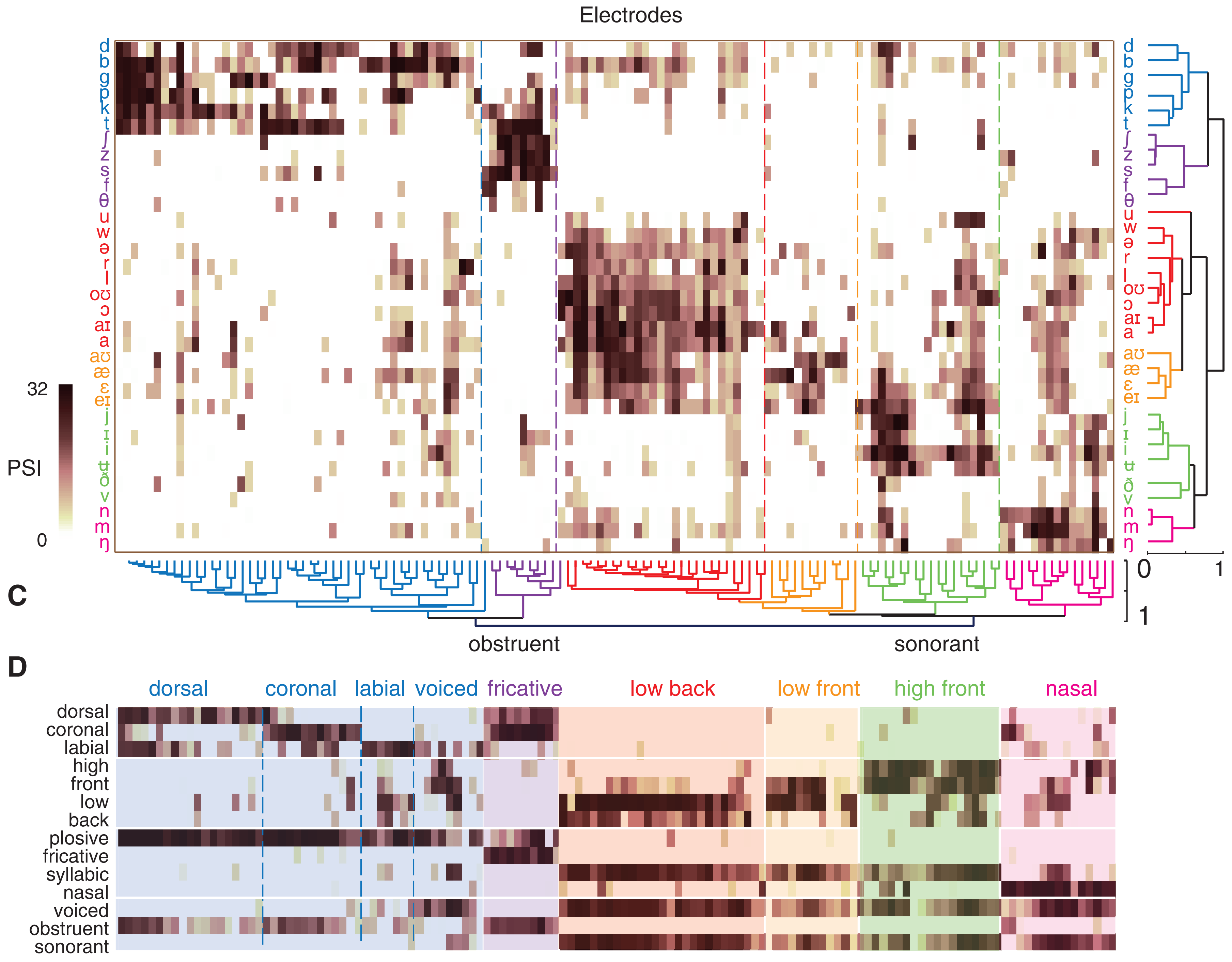
EVIDENCE FOR SENSITIVITY TO PHONETIC FEATURES

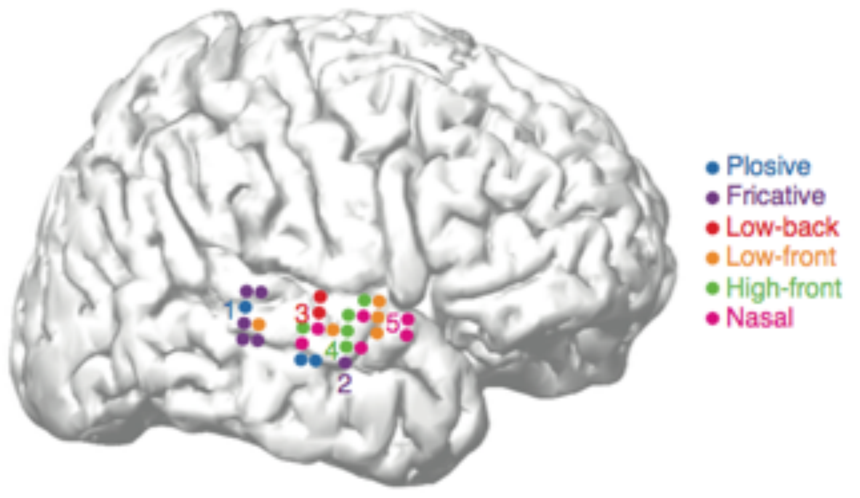






Chang et al. (2010)
Nature Neuroscience



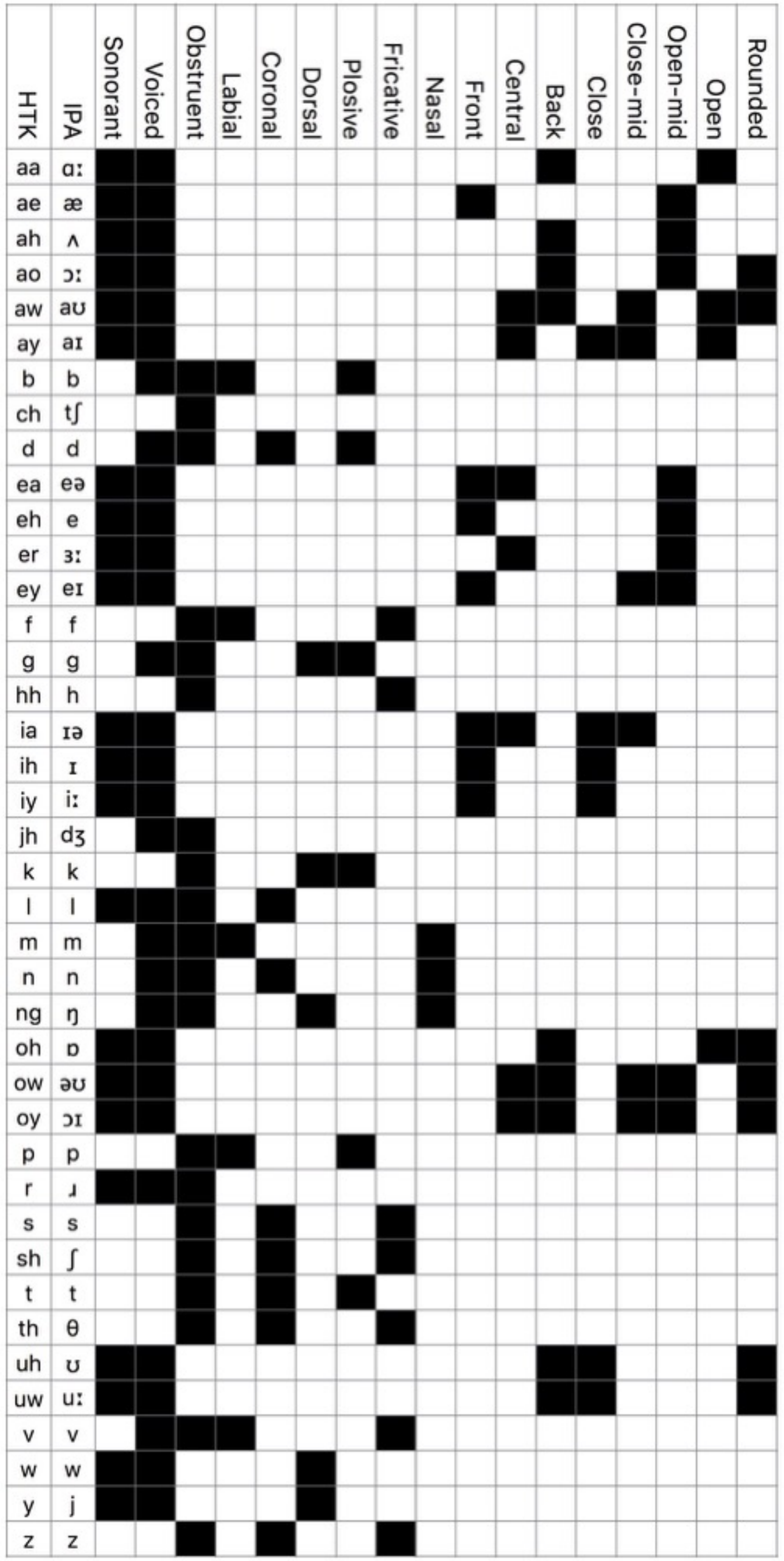


Mesgarani et al. (2014)
Science

MAPPING PHONETIC FEATURES

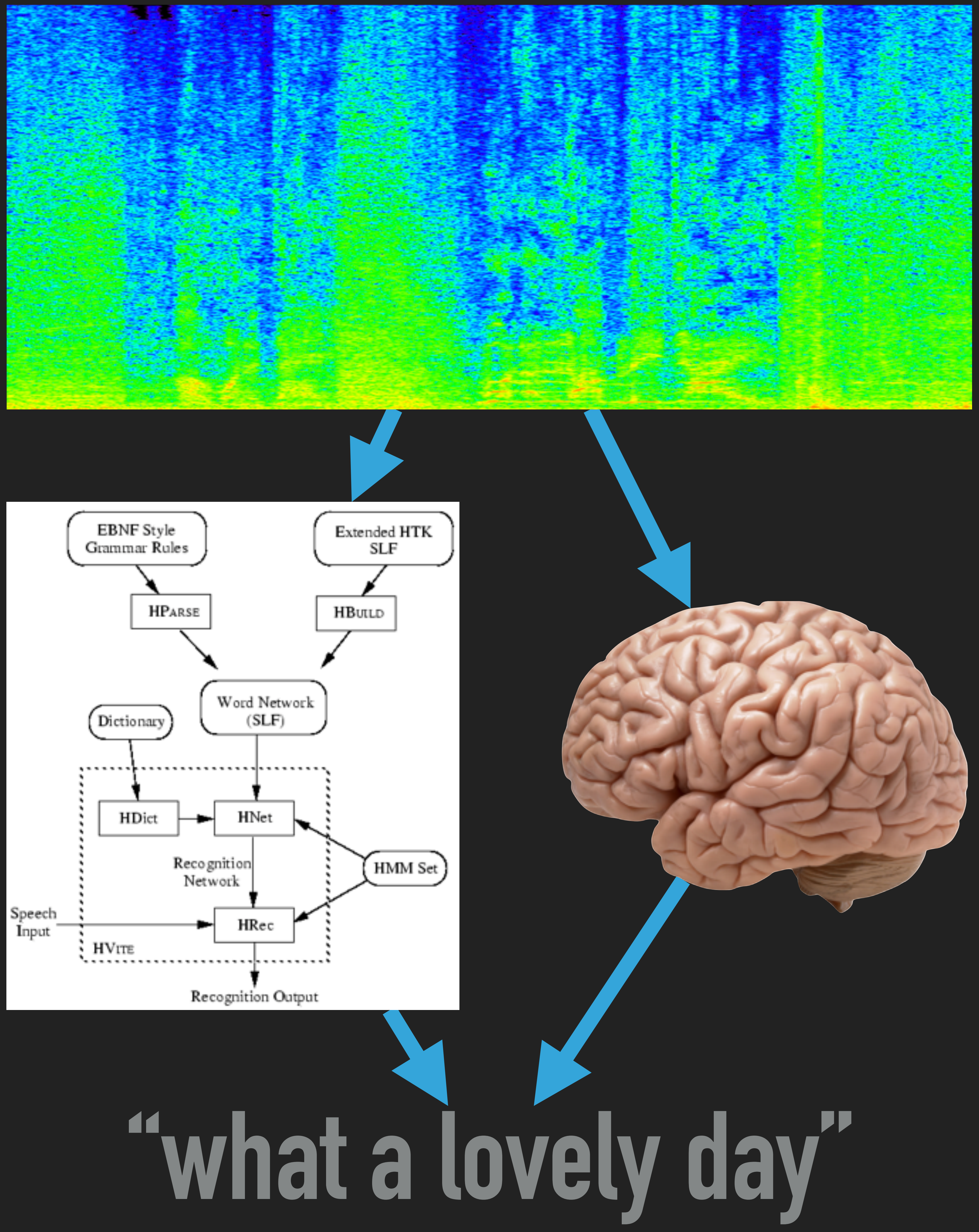
- We aim to investigate neural representations of the phonetic features in our stimuli.
- 400 English words presented aurally to native English speakers in a MEG scanner with simultaneous EEG.
  - (Data collected for another experiment, so not perfectly suited.)
- Same words presented to ASR system.

Broad categories	Sonorant
	Voiced
	Obstruent
Place	Labial
	Coronal
	Dorsal
Manner	Plosive
	Fricative
	Sibilant
	Nasal
Vowel frontness	Front
	Central
	Back
Vowel closeness	Close
	Close-mid
	Open-mid
	Open
	Rounded



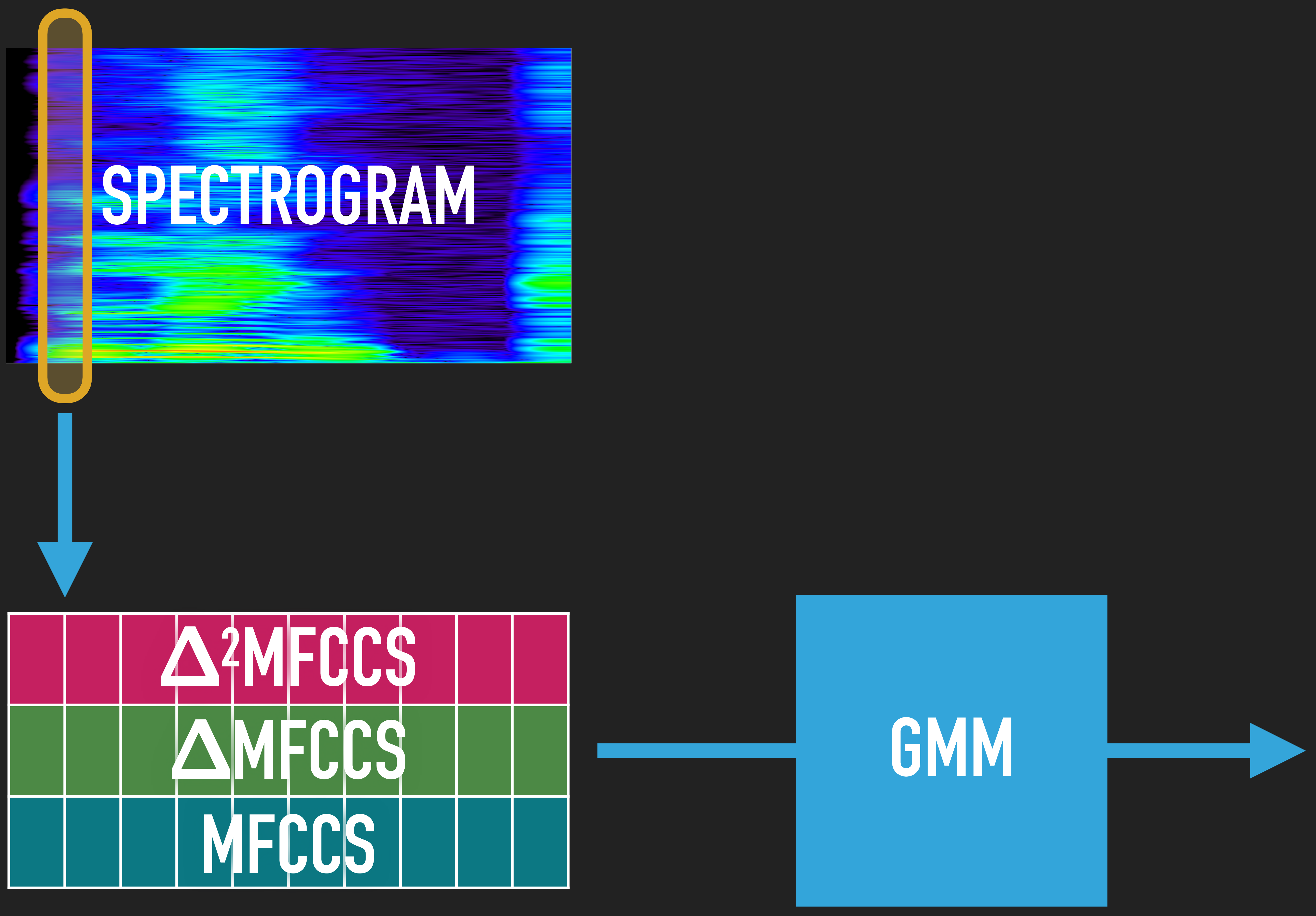
AUTOMATIC SPEECH RECOGNISERS

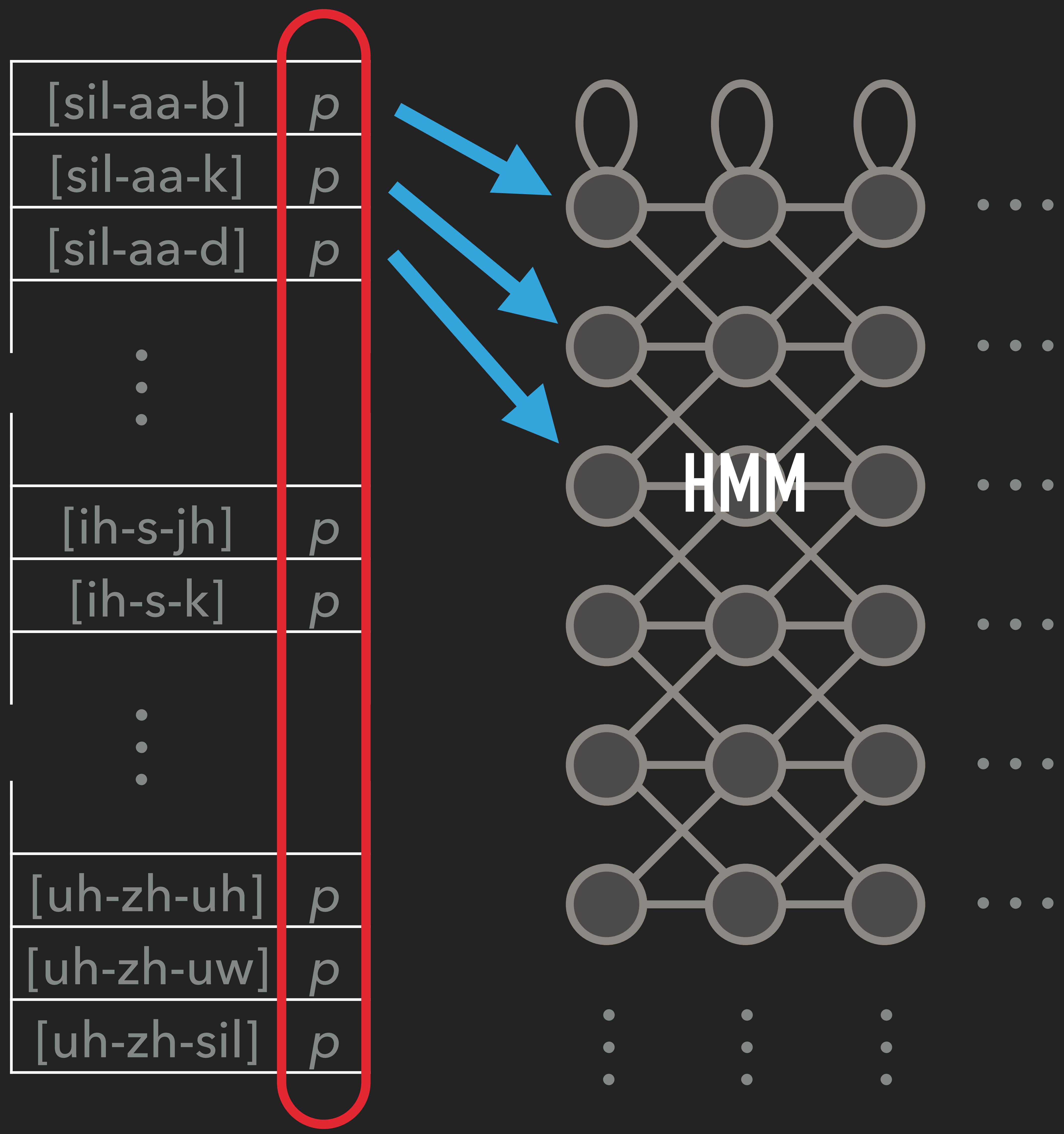
- Automatic speech recognisers perform (part of) the same task as humans.
- Unlike in some models of computer vision, most ASR systems aren't architecturally inspired by biological systems.
  - Partially because only humans understand speech.
- We "reverse-engineer the engineering solution" to model phonetic content of our spoken language.



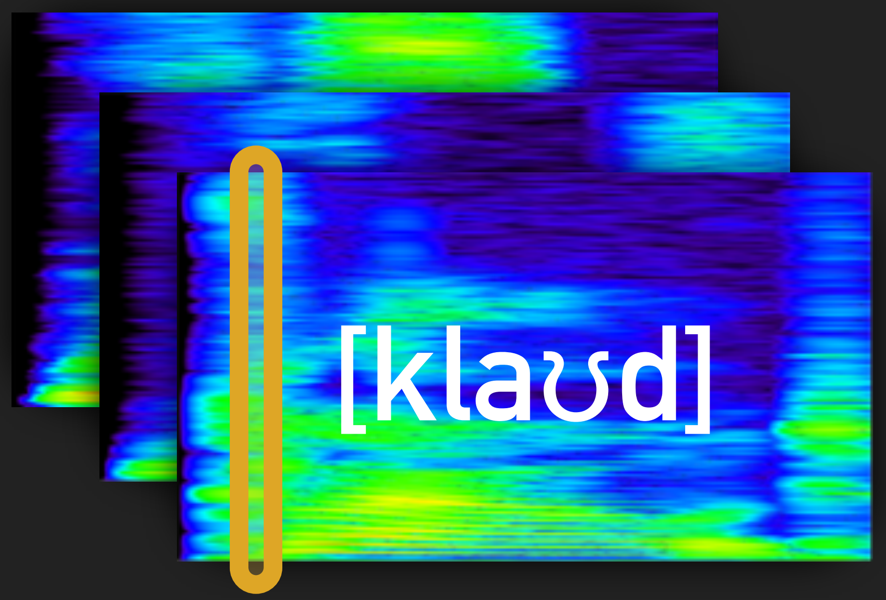
Young et al. (1997)
The HTK Book

HTK: HIDDEN MARKOV MODEL TOOLKIT





PHONETIC RDMS



Every frame
(10ms)

Every word
(400)

[aa]

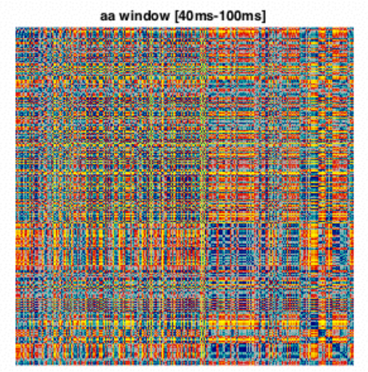
Every phone
(40)

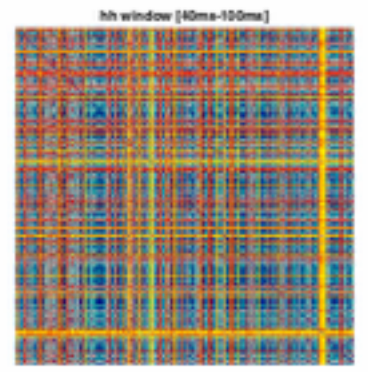
[sil-aa-sil]	*p*	*p*	*p*	
[sil-aa-b]	*p*	*p*	*p*	• • •
[sil-aa-k]	*p*	*p*	*p*	
⋮				
[z-aa-v]	*p*	*p*	*p*	
[z-aa-z]	*p*	*p*	*p*	• • •
[z-aa-sil]	*p*	*p*	*p*	

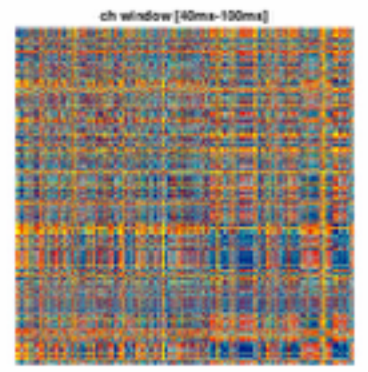
Every
triphone

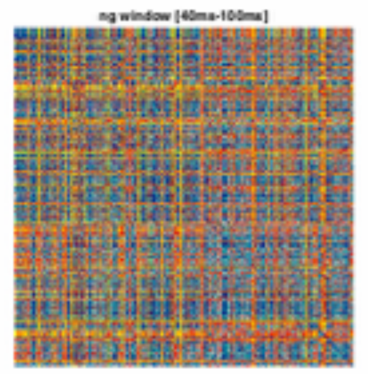
Sliding window

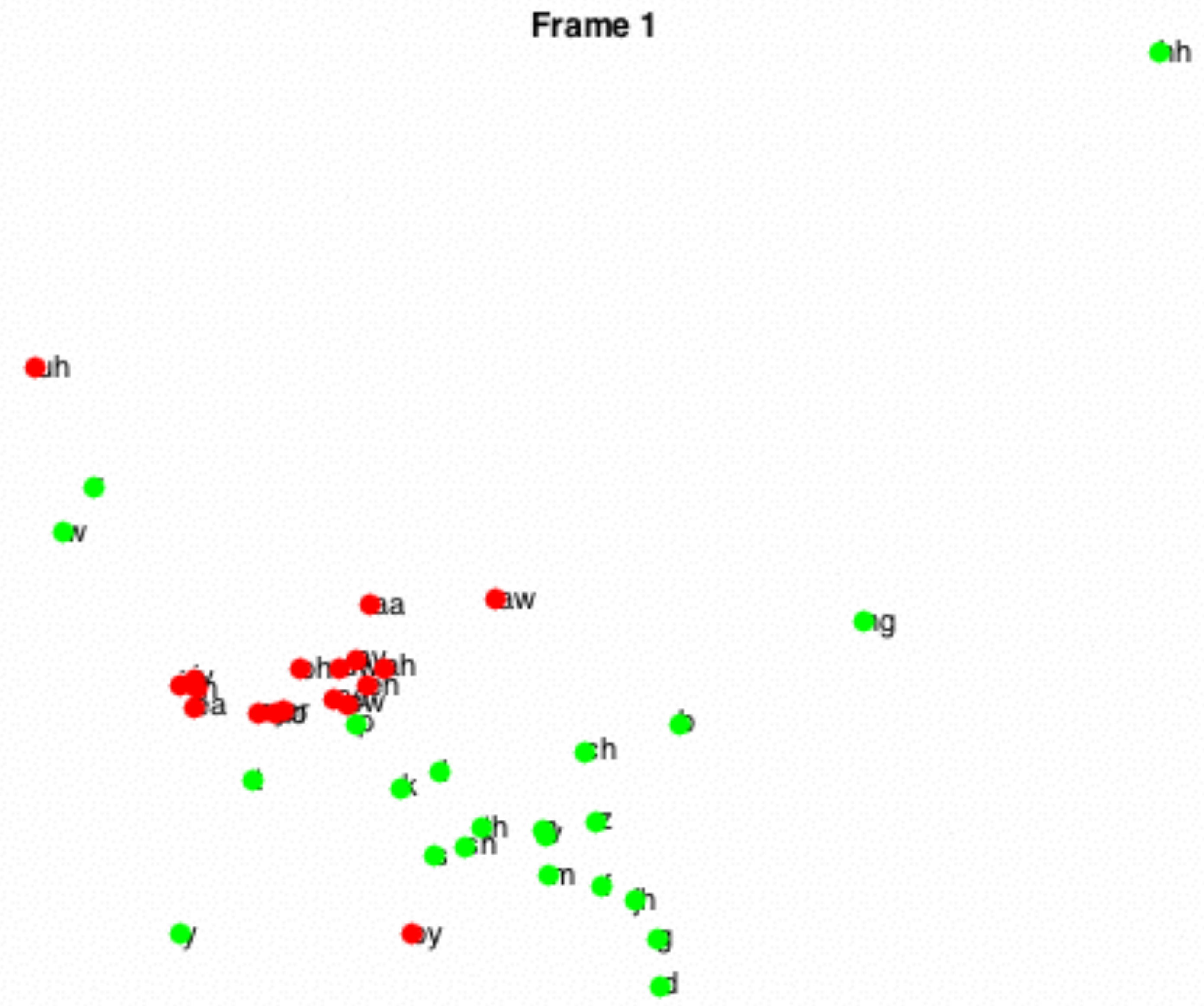
MODEL RDM STRUCTURE

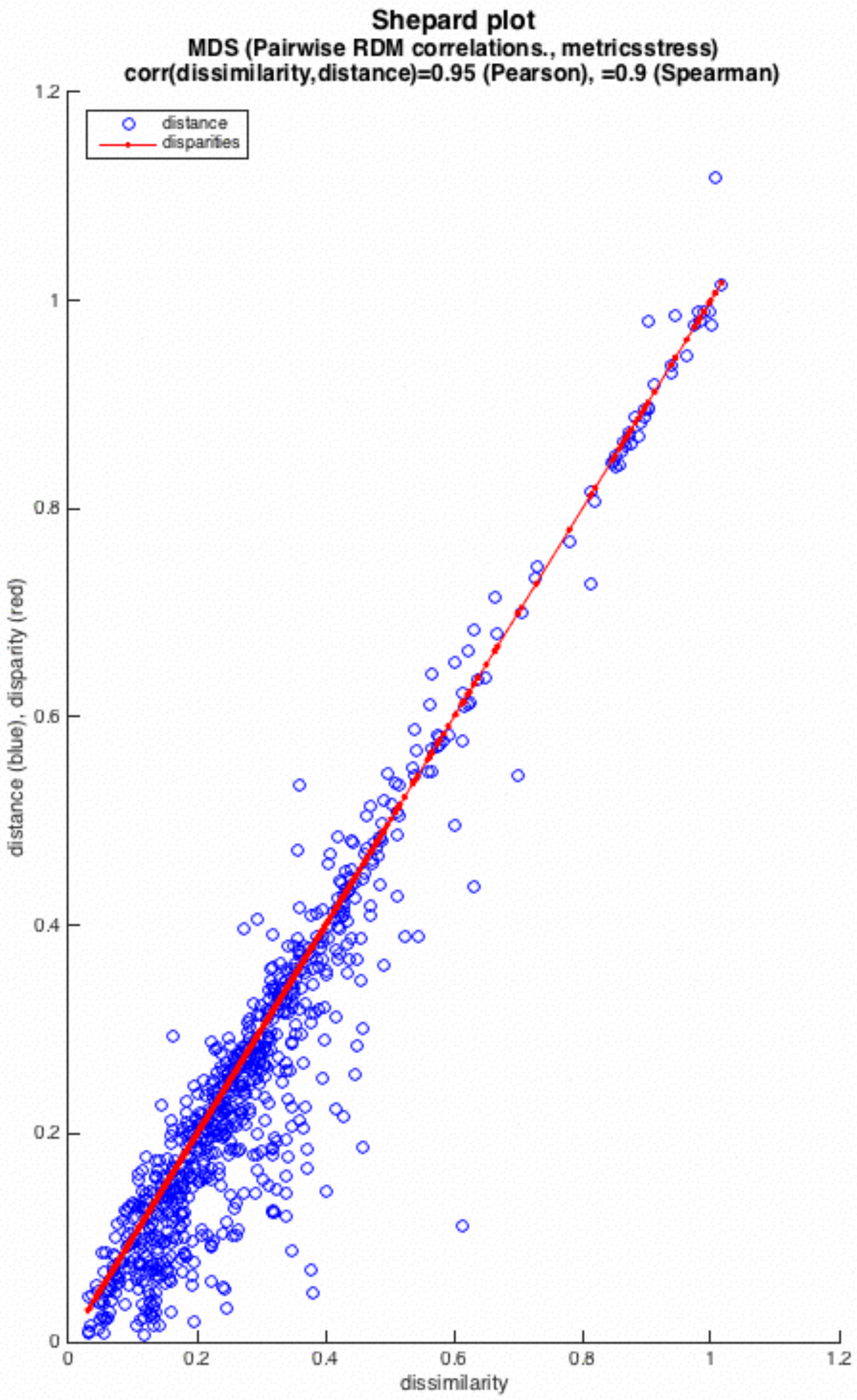




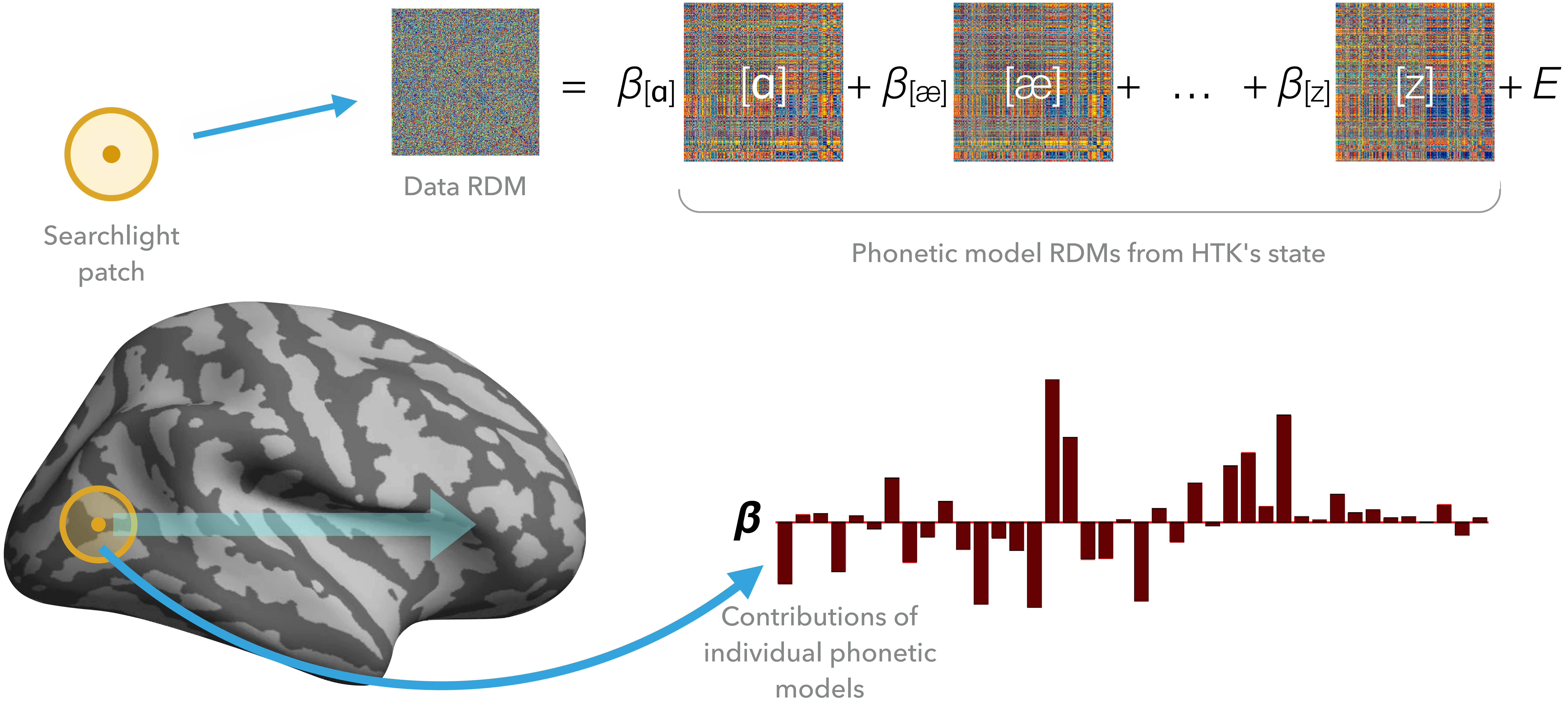




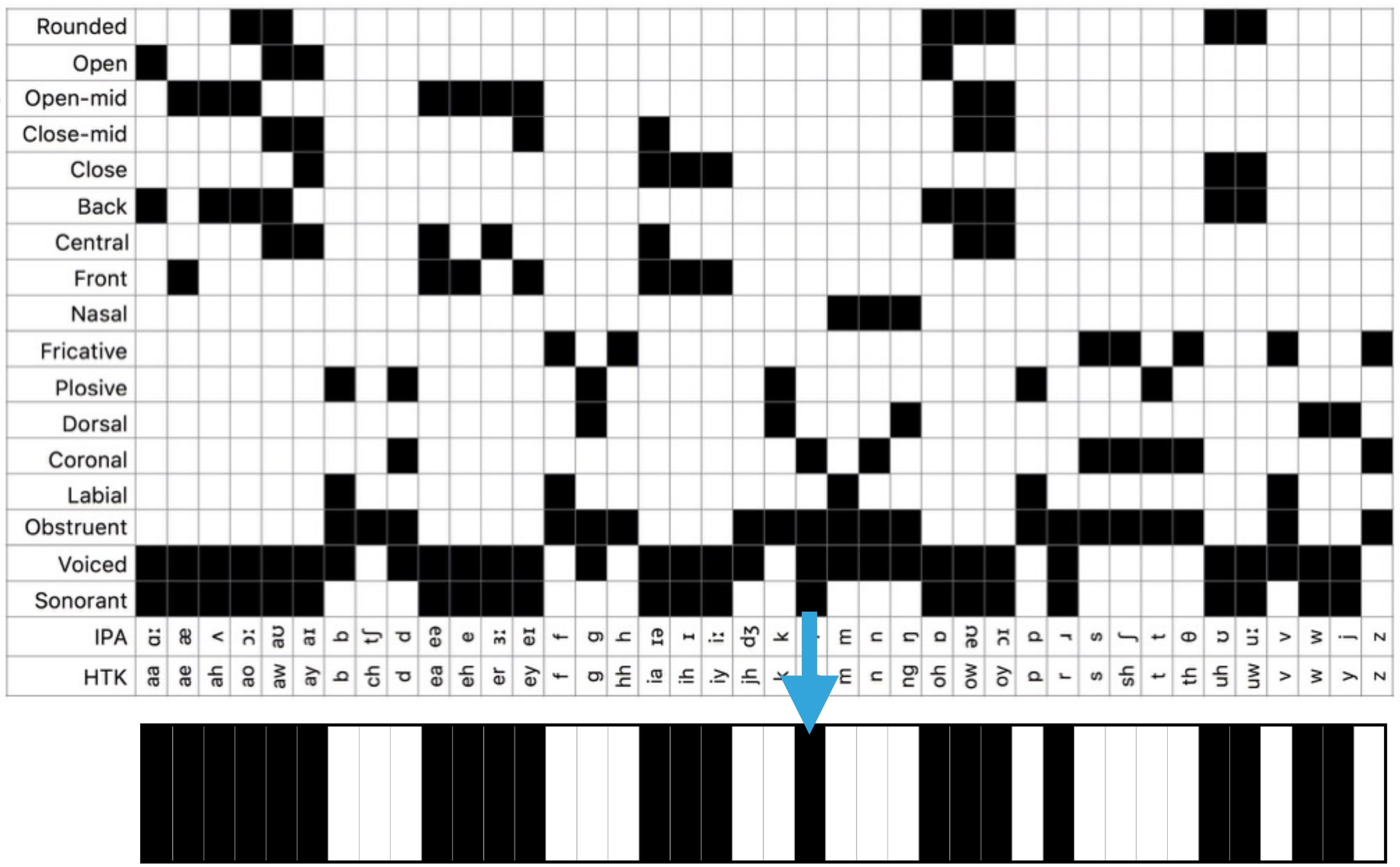




SEARCHLIGHT ANALYSIS

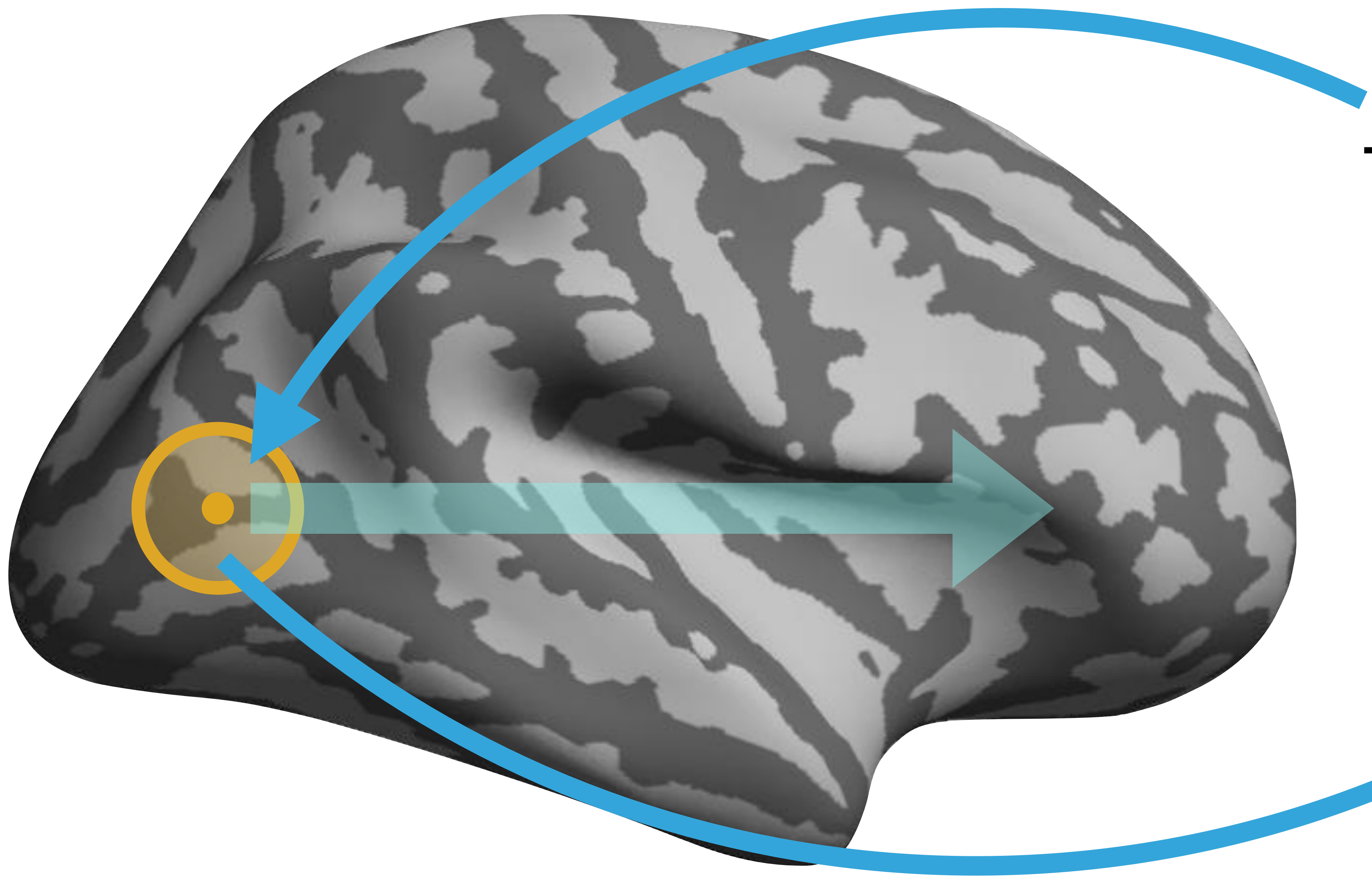


SEARCHLIGHT ANALYSIS



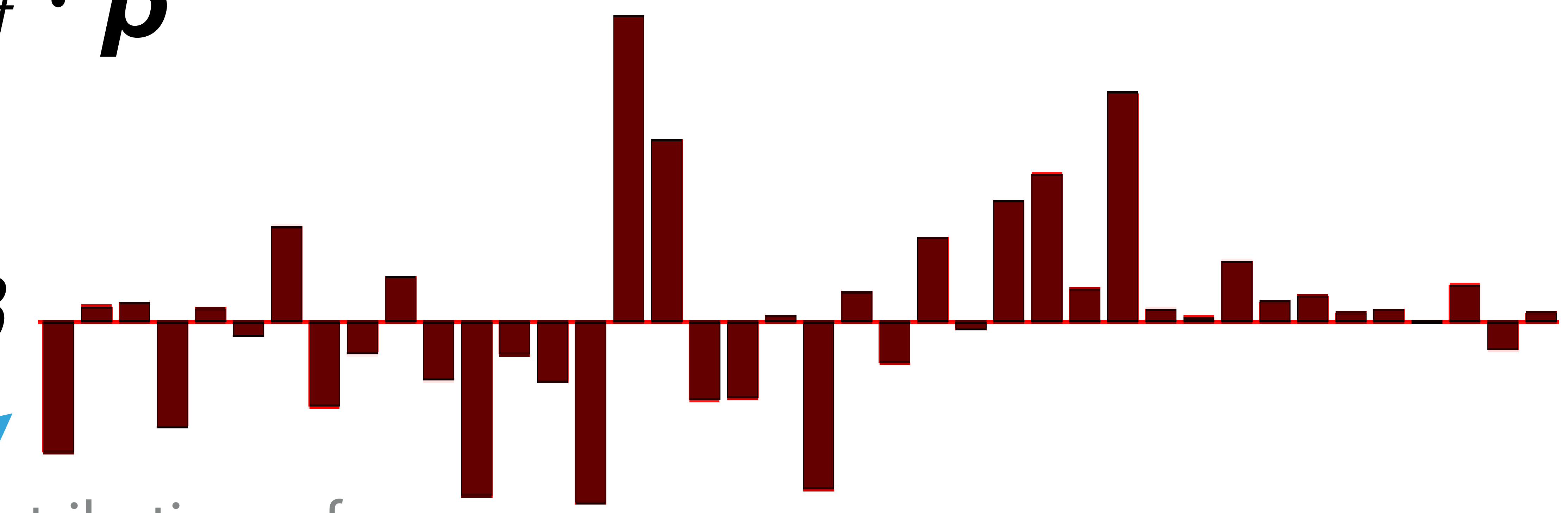
$\boldsymbol{X}$sonorant

Map of
"sonorant" feature



$$\text{fit}_f = \boldsymbol{X}_f \cdot \boldsymbol{\beta}$$

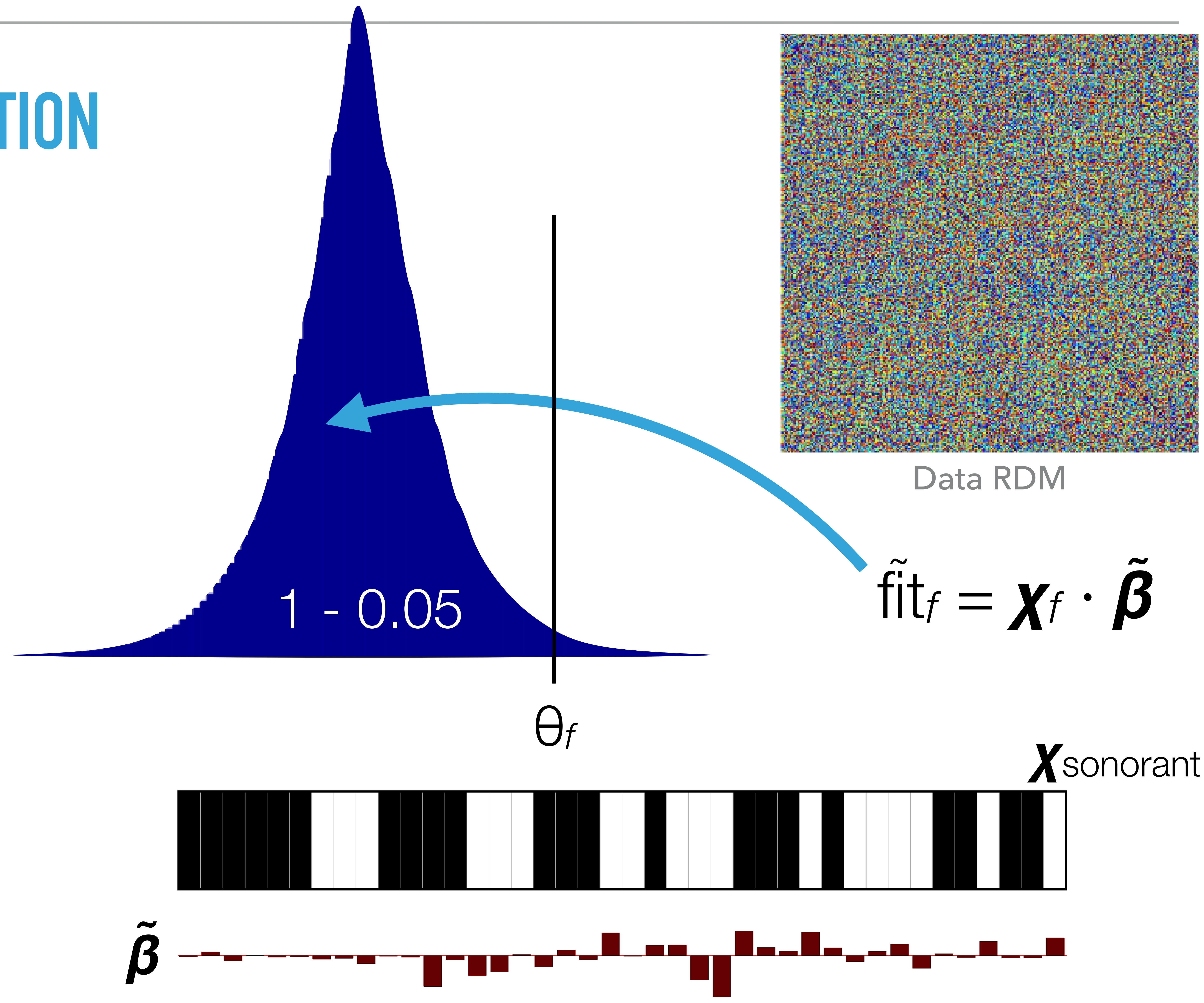
$\boldsymbol{\beta}$



Contributions of
individual phonetic models

SIMULATING THE NULL DISTRIBUTION

- Under the null hypothesis, there is no difference between experimental conditions.
- So, we can permute word labels (rows and columns of data RDM) and would expect no difference in fit.
- Aggregate 1000s of fits from randomly permuted data RDMs.
  - This our simulated null distribution.
- We threshold our maps of fit with θ_f.

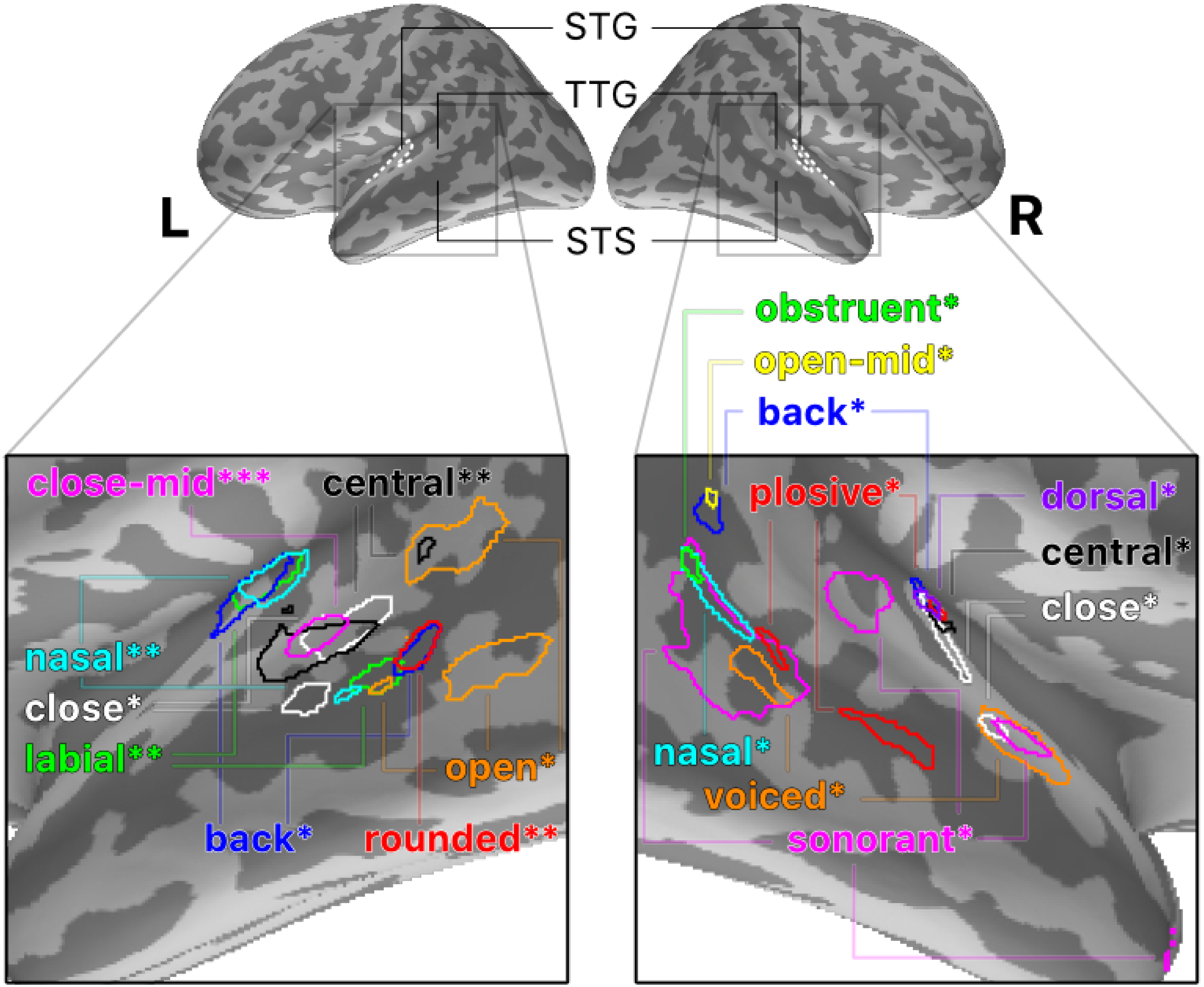


RESULTS

REGIONS OF PHONETIC FEATURE REPRESENTATION

RESULTS

- Most (not every) feature we tested showed super-threshold fit in and around auditory cortex.
- Features describing broad categories fit best on the right.
- Regions of fit on the left tended to be more focussed.
- Within-category features showed fits bilaterally.



SUMMARY

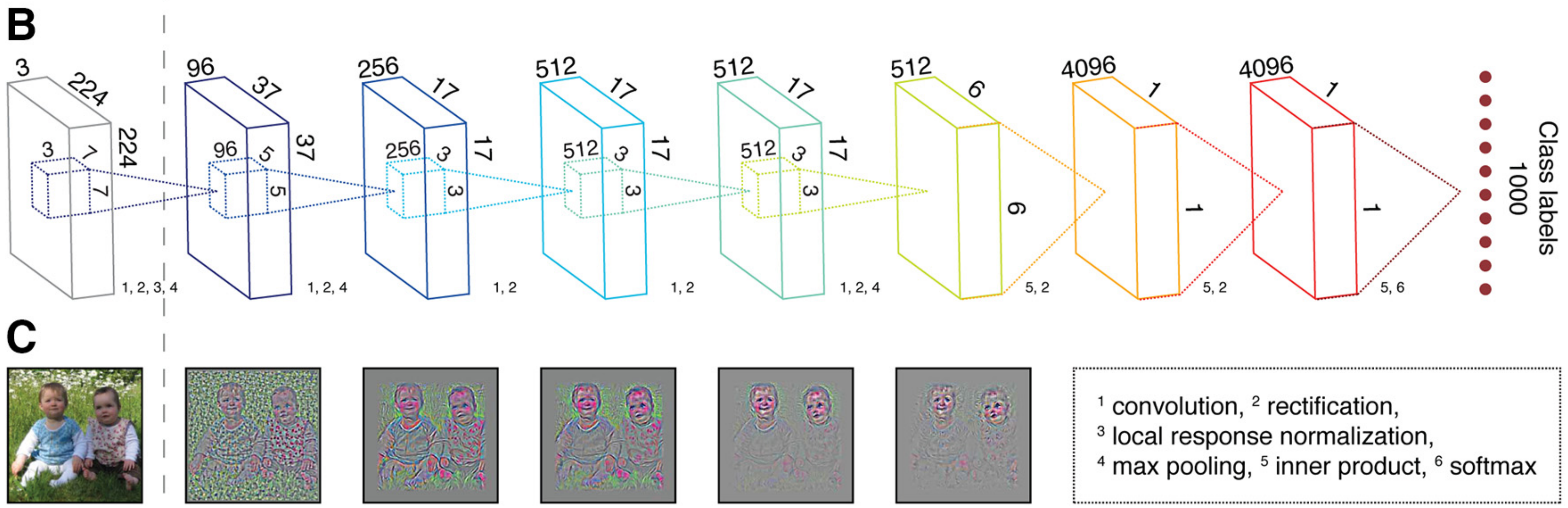
- Evidence of regions of phonetic feature sensitivity in human auditory cortex.
- Use multivariate pattern analysis methods (cf. classical contrasts) to understand individual representations.
- Model features relevant to speech comprehension using machine ASR systems.
- RSA allows comparison of brain states and machine states.
- EMEG records rich brain response data, non-invasively.
- Early sound-to-meaning mappings are still poorly understood.

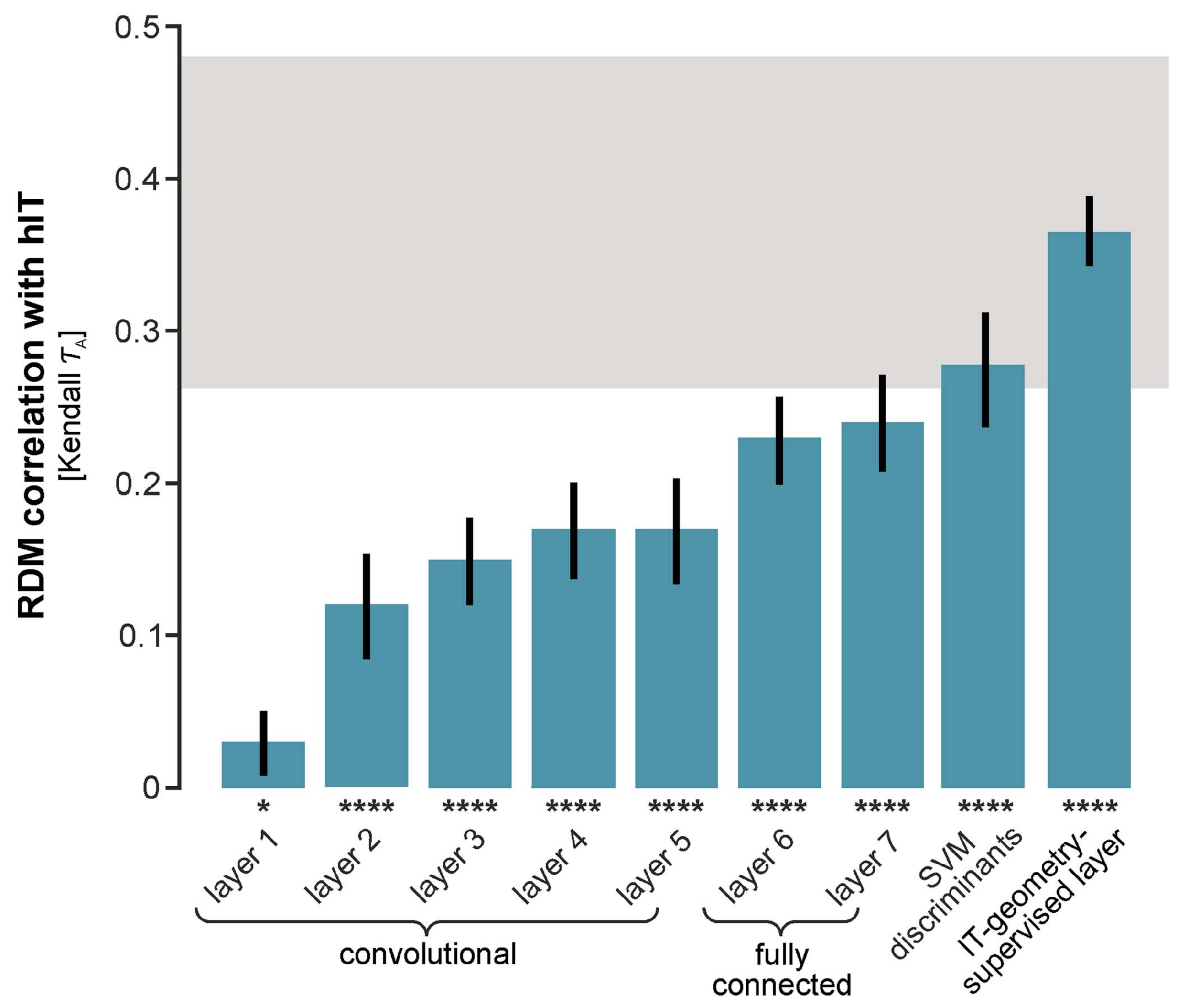
WHERE NEXT?

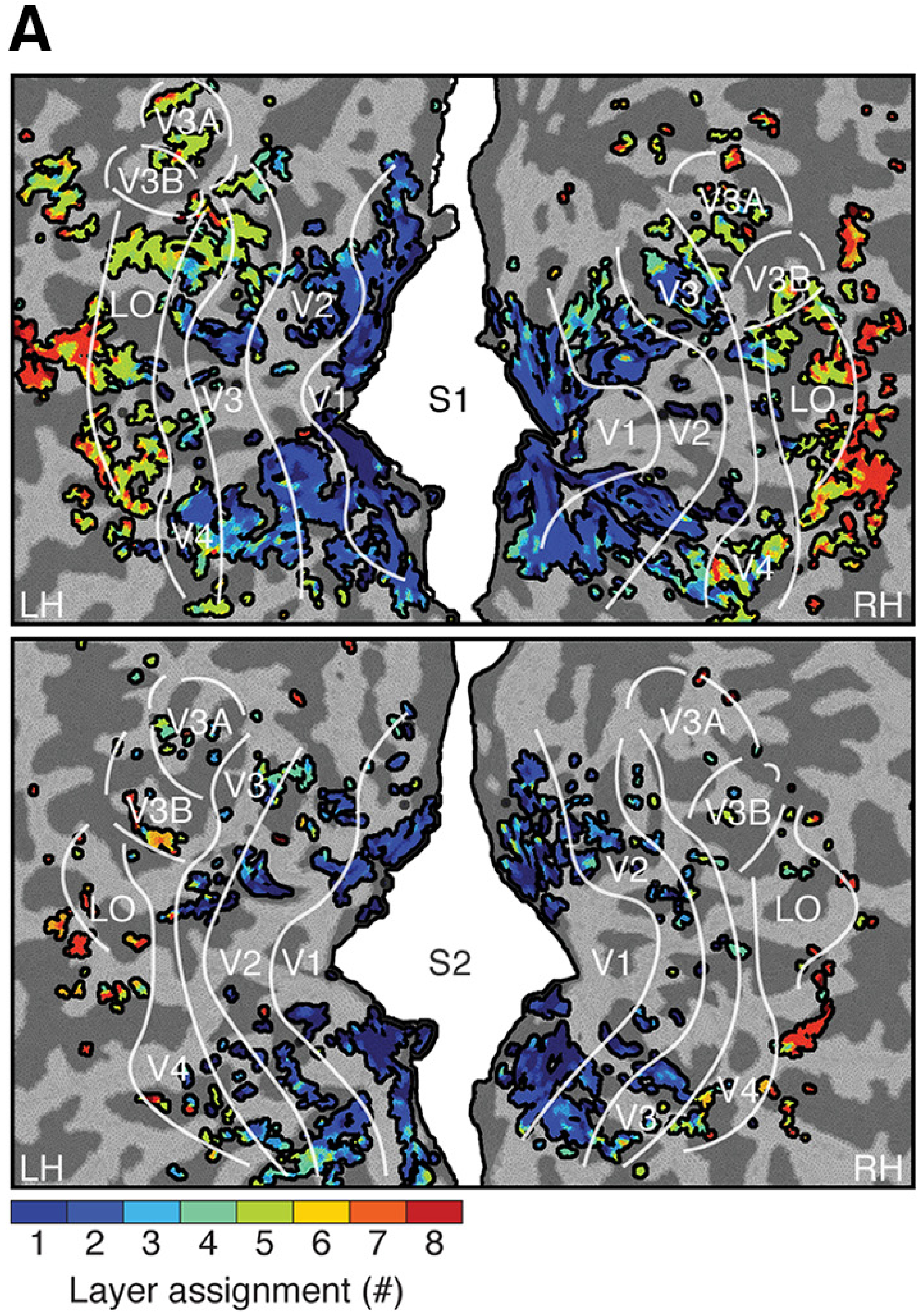
- Use a deep artificial neural network-based ASR.
- Don't rely on phone-level representation.
    - Use "bottom-up" features.
    - Hidden-layer representations.
- Understand time-resolved results.
- Better data.
    - Continuous speech.
- Next level: semantics from abstract labels.

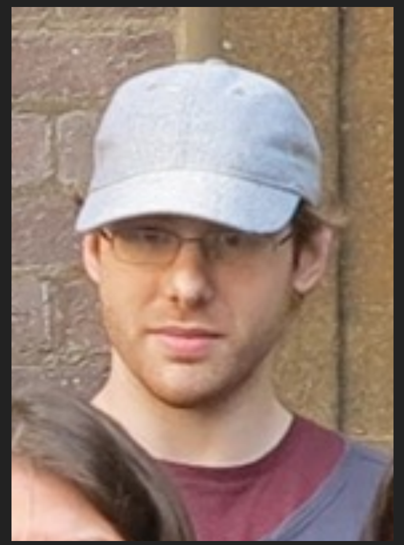
Khaligh-Razavi & Kriegeskorte (2014)
PLOS Computational Biology

Güçlü & Gerven (2015)
Journal of Neuroscience








Andrew Thwaites


Elisabeth Fonteneau


Cai Wingfield


William Marslen-Wilson

Department of Psychology


Xunying Liu


Chao Zhang


Phil Woodland

Department of Engineering


Li Su

Department of Psychiatry