Cai Wingfield

CSLB, Department of Psychology
University of Cambridge

Workshop on Neurocomputation: From Brains to Machines
25 November 2015

Understanding human speech recognition: Reverse-engineering the engineering solution using EMEG and RSA

Brains and Machines

- We've seen from previous speakers how:
  - Machine systems are designed to perform the same tasks as humans.
  - The architecture of machine models of (e.g.) vision may relate to those of biological systems.
  - By using methods such as RSA, intermediate-level derived representations in one may be compared to those in the other.

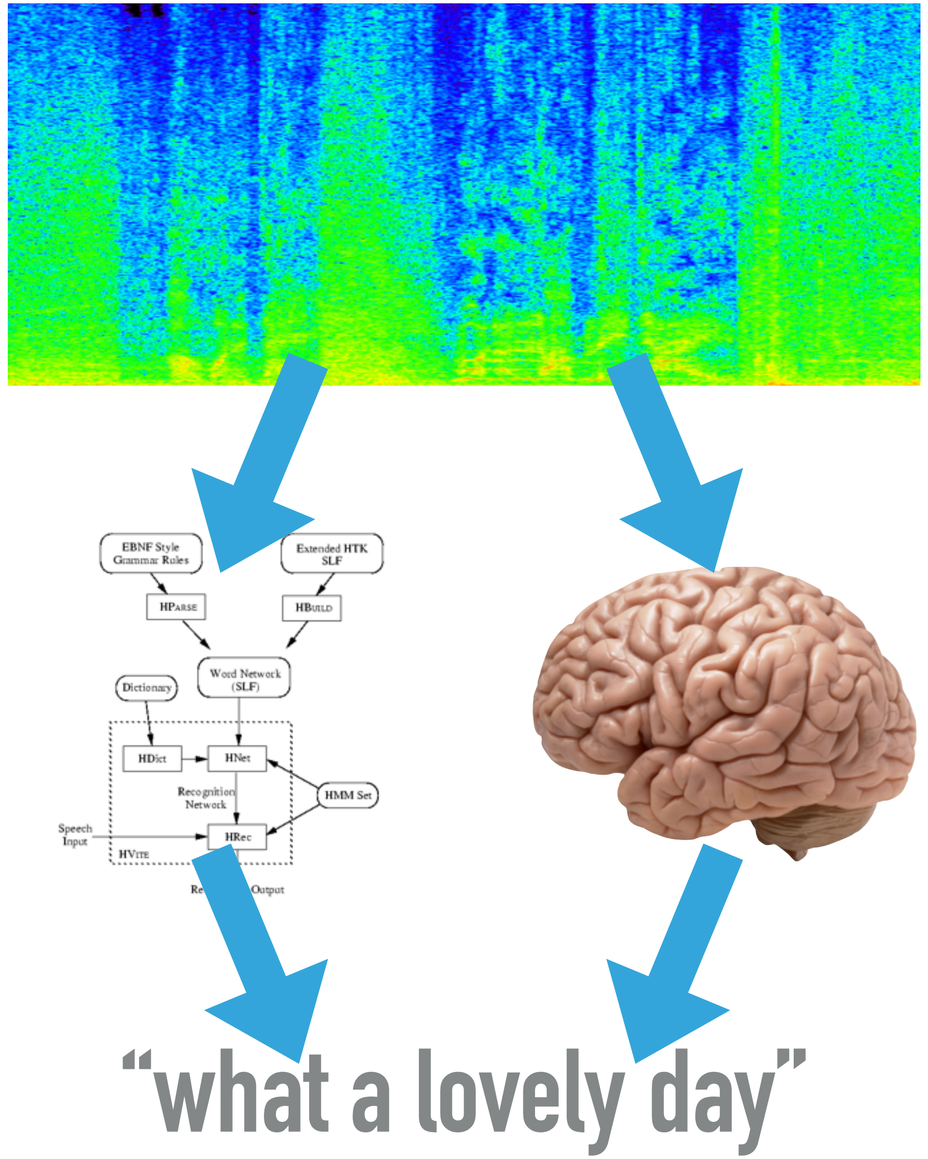
Speech and vision

- Unlike visual objects, speech stimuli are time-sensitive.
- There's no standard neurocomputational model of speech comprehension.
  - Humans alone amongst animals have this faculty.
- The most effective artificial systems' designs don't tend to relate to biological models.

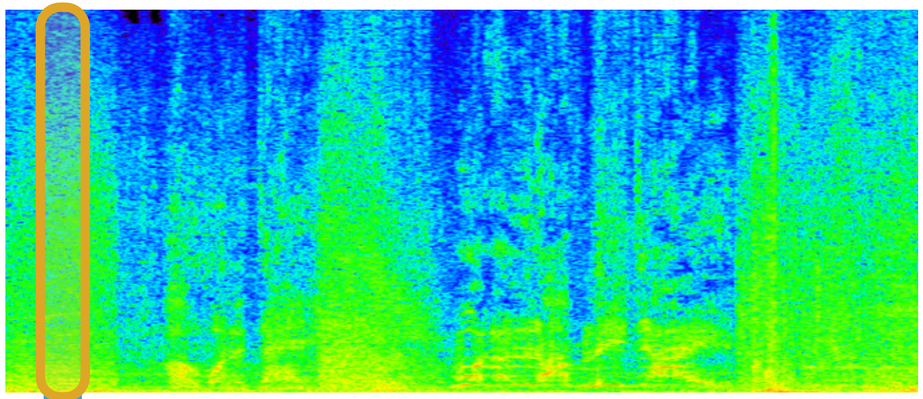
- However, machines provide *a* computational model of the process.

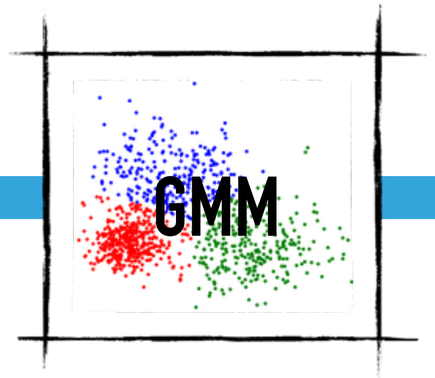
Speech recognition

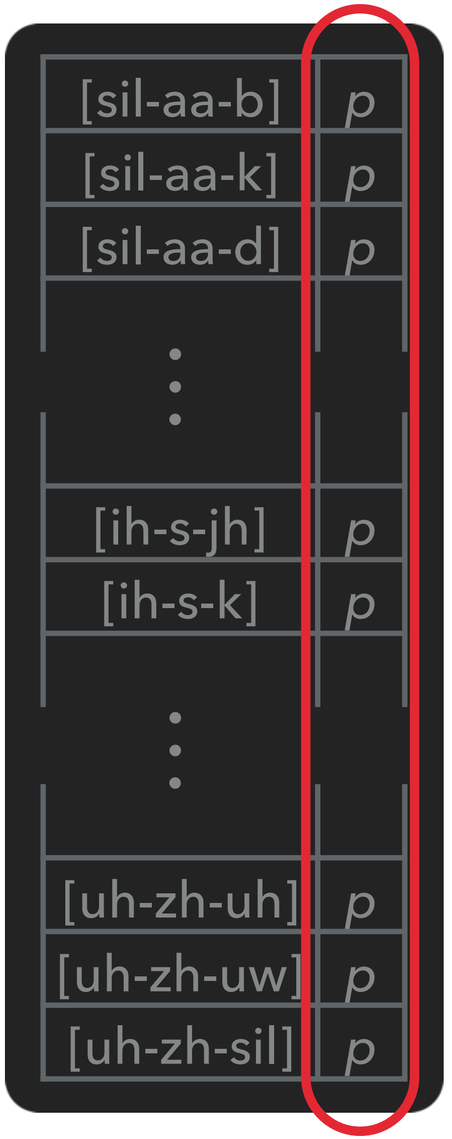
- Both human brains and machines can recognise speech accurately.
  - Transforms raw acoustic input into abstract word “objects”.
  - Artificial (ASR) systems are nearly as good as humans.
- In brains, this is mediated by some complex, poorly understood neurobiological process.
- We will compare intermediate-level representations in an ASR and human auditory cortex using RSA.

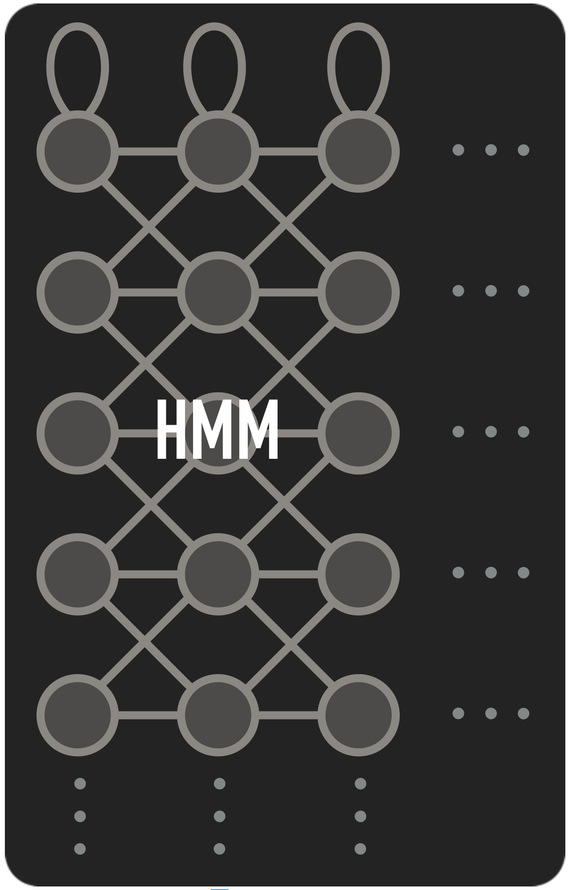


HTK: GMM-HMM







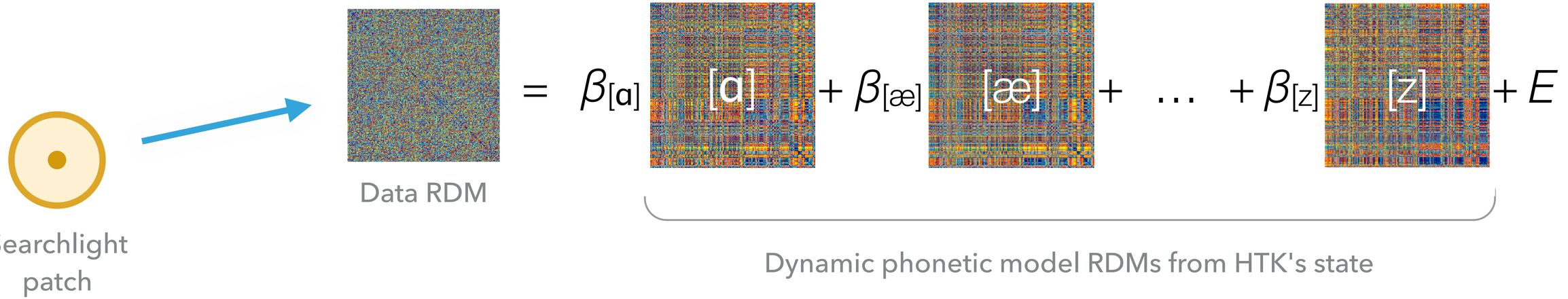


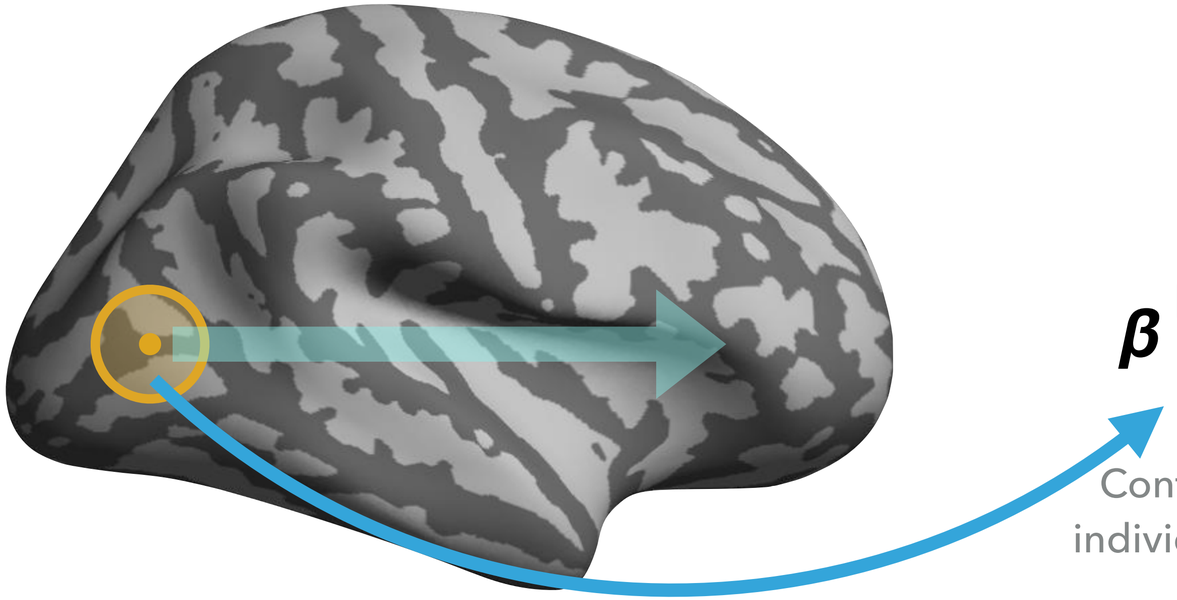
[sil-w-oh] [w-oh-t] [oh-t-sil]

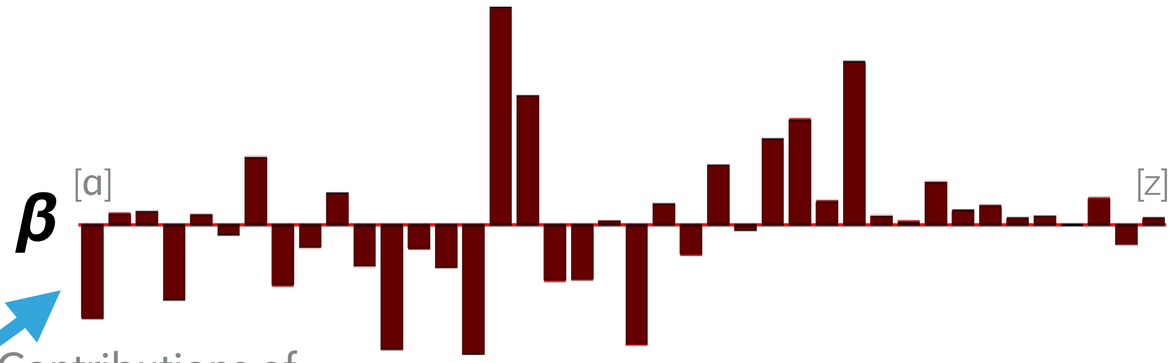
WHAT

Young et al. (1997)
The HTK Book

Searchlight GLM RSA



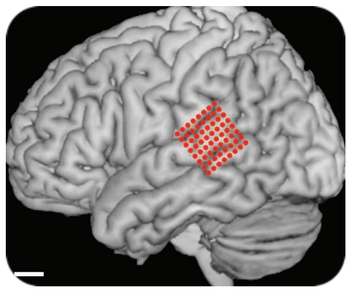


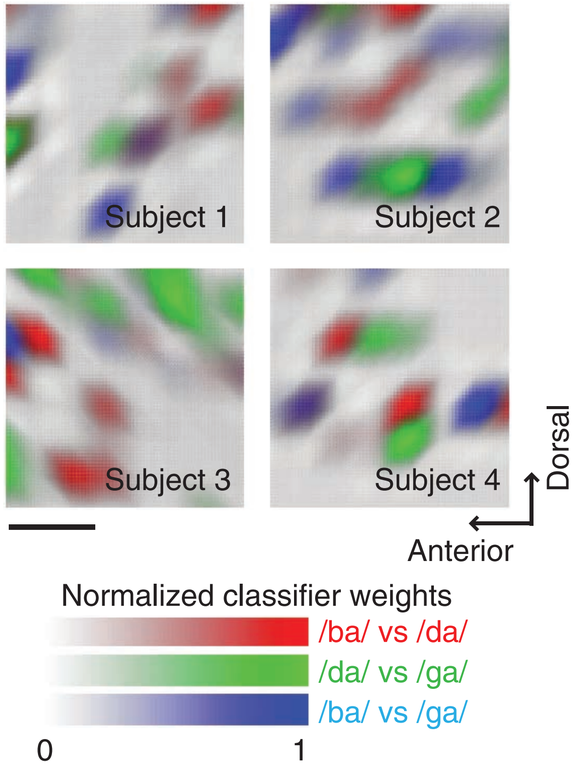


Contributions of individual phonetic models

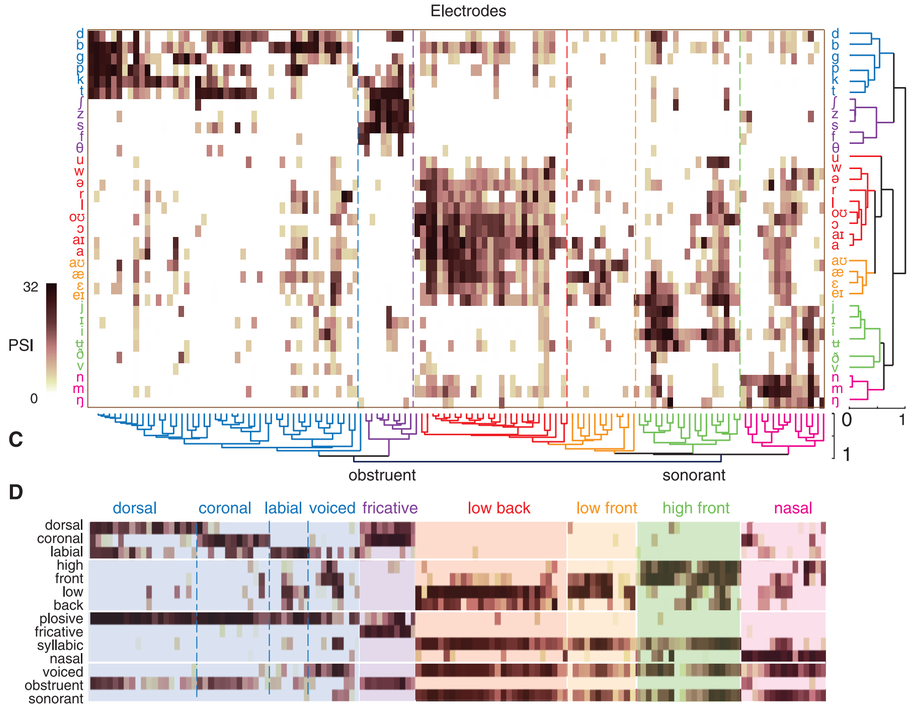
Su et al. (2014)
Frontiers in Neuroscience

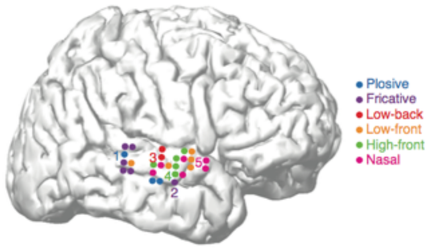
Evidence for sensitivity to phonetic features





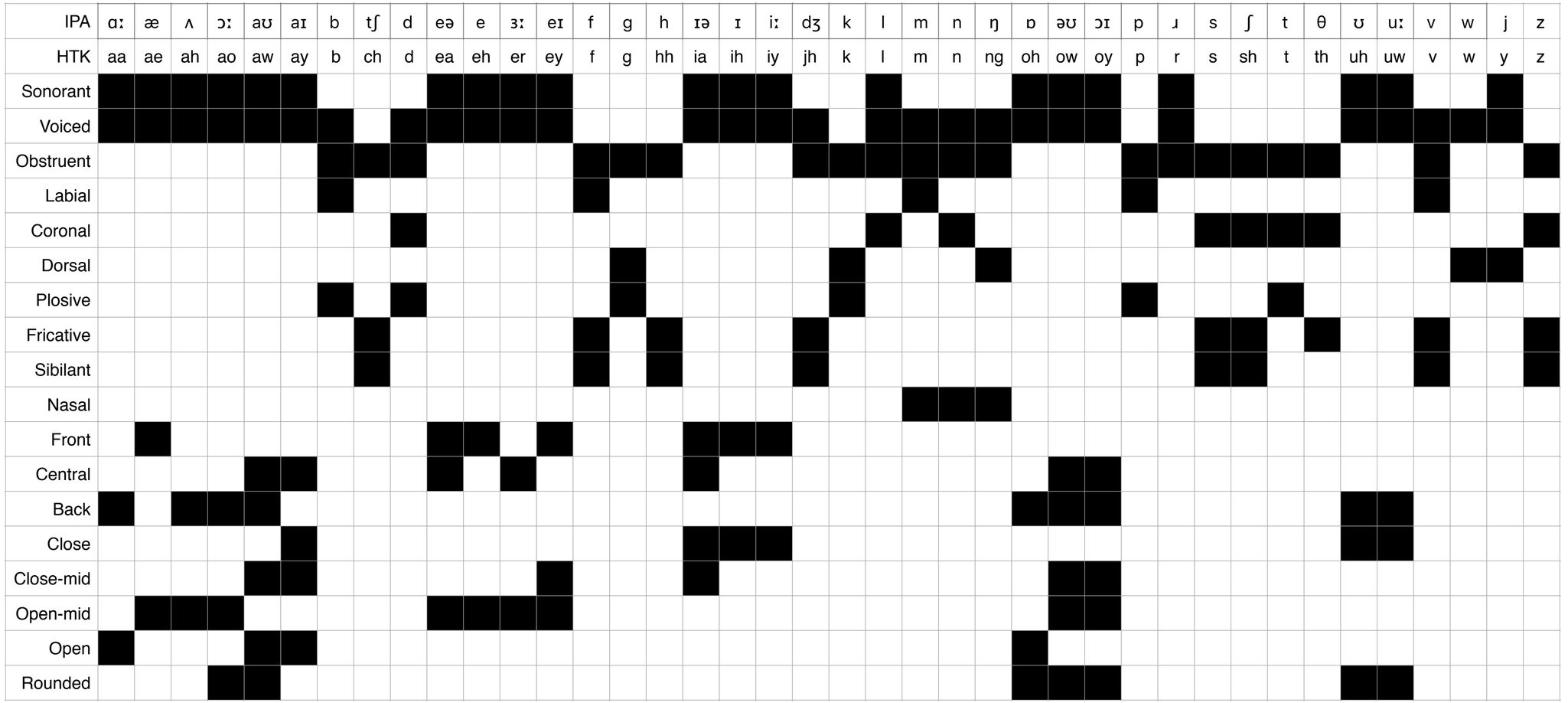
Chang et al. (2010)
Nature Neuroscience





Mesgarani et al. (2014)
Science

Evidence for sensitivity to phonetic features



Broad categories: Sonorant, Voiced, Obstruent

Place: Labial, Coronal, Dorsal

Manner: Plosive, Fricative, Sibilant, Nasal

Vowel frontness: Front, Central, Back

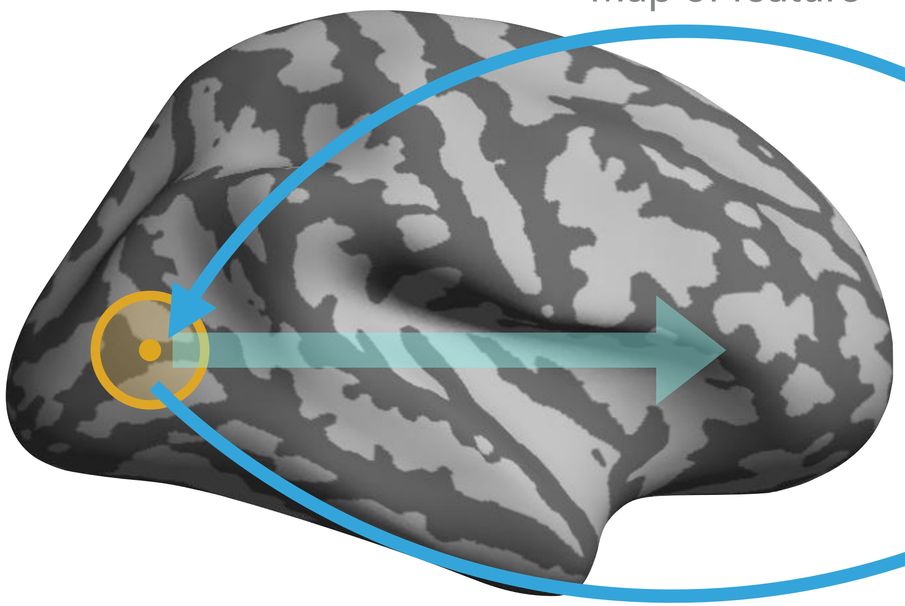
Vowel closeness: Close, Close-mid, Open-mid, Open

Rounded

Searchlight GLM RSA

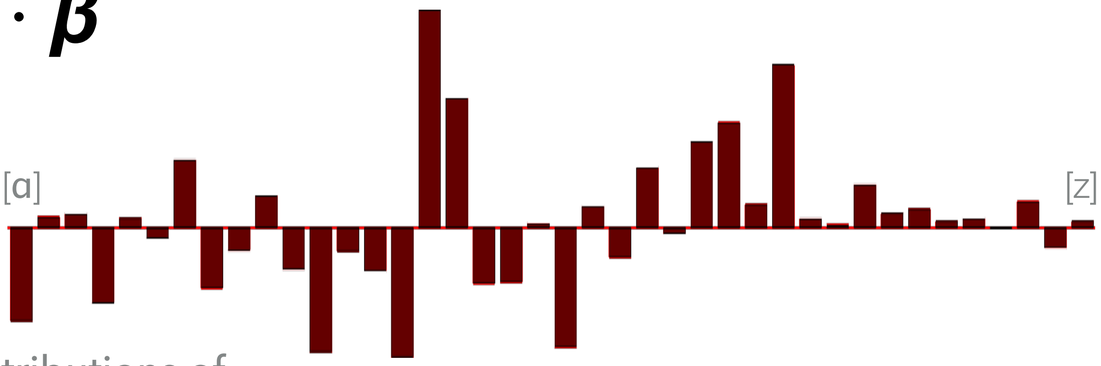
	IPA	ɑː	æ	ʌ	ɔː	aʊ	aɪ	b	tʃ	d	eə	e	ɜː	eɪ	f	g	h	ɪə	ɪ	iː	dʒ	k	l	m	n	ŋ	ɒ	əʊ	ɔɪ	p	ɹ	s	ʃ	t	θ	ʊ	uː	v	w	j	z
	HTK	aa	ae	ah	ao	aw	ay	b	ch	d	ea	eh	er	ey	f	g	hh	ia	ih	iy	jh	k	l	m	n	ng	oh	ow	oy	p	r	s	sh	t	th	uh	uw	v	w	y	z
Sonorant		■	■	■	■	■	■				■	■	■	■				■	■	■			■				■	■	■		■					■	■			■	
Voiced		■	■	■	■	■	■	■		■	■	■	■	■				■	■	■	■		■	■	■		■	■	■		■					■	■	■	■	■	
Obstruent								■	■	■					■	■	■				■	■	■	■	■					■	■	■	■	■	■			■			■
Labial								■							■								■							■								■			
Coronal										■													■		■							■	■	■	■						■
Dorsal																■						■				■												■	■		
Plosive								■		■						■						■								■				■							
Fricative									■					■		■	■															■	■		■			■			■
Sibilant									■					■		■																■	■					■			■
Nasal																								■	■																
Front		■									■	■		■				■	■	■																					
Central						■	■				■		■					■											■	■												
Back		■		■	■	■																					■	■	■								■	■			
Close							■												■	■																		■	■			
Close-mid						■	■							■				■	■										■	■												
Open-mid			■	■	■						■	■	■	■															■	■												
Open		■				■	■																				■														
Rounded					■	■																					■	■	■								■	■			

Map of feature



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

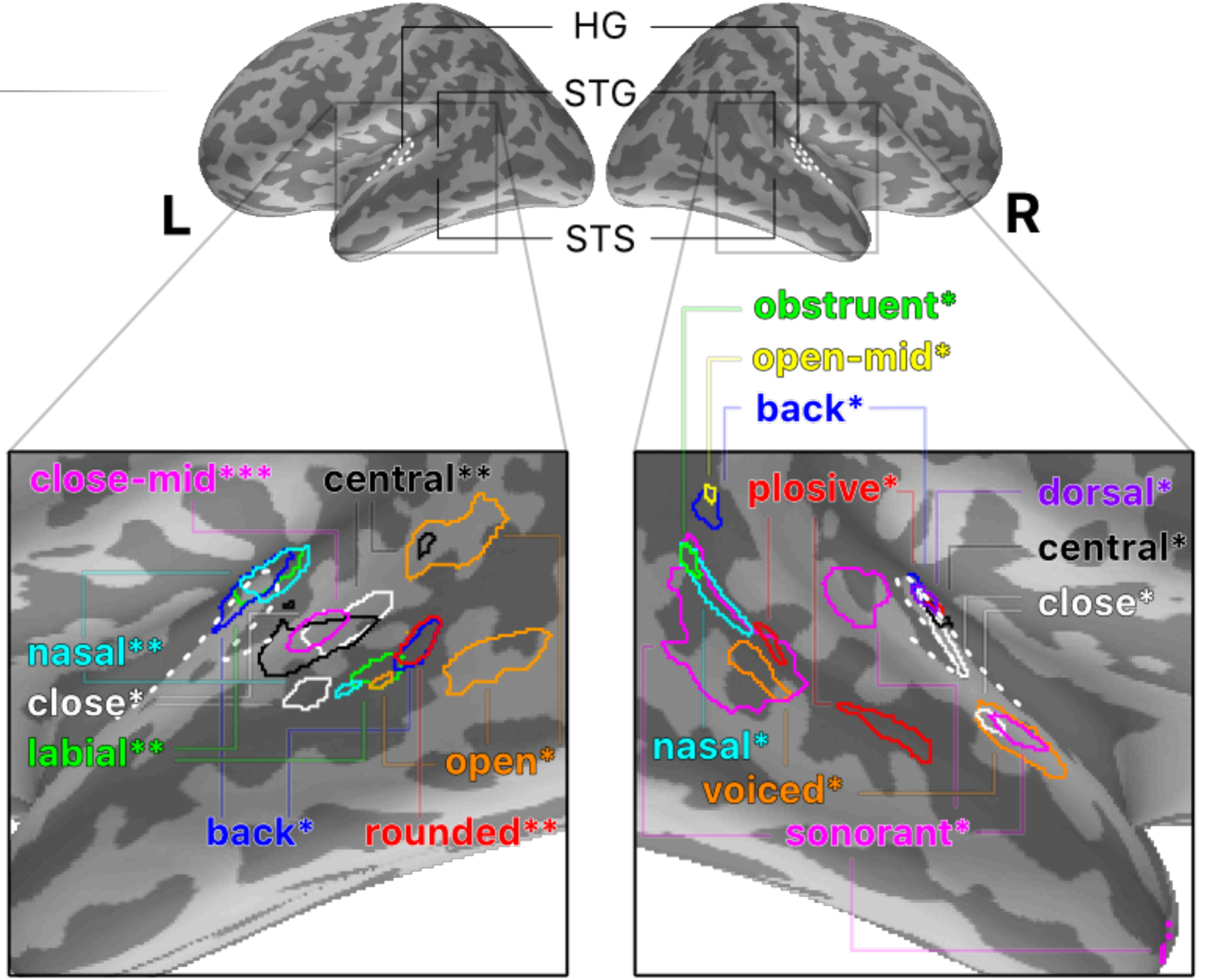
$\boldsymbol{\beta}$ [ɑ] [z]



Contributions of individual phonetic models

Speech recognition

- 16 subjects, 400 words, EMEG.
- Most features we tested showed significant fit in auditory cortex.
  - Bilateral HG, STG, STS.
- Broad category features fit best on the right.
- Regions on the left tended to be more focussed.
- Within-category features showed fits bilaterally.



[100, 170] ms

Wingfield et al. (in prep.)

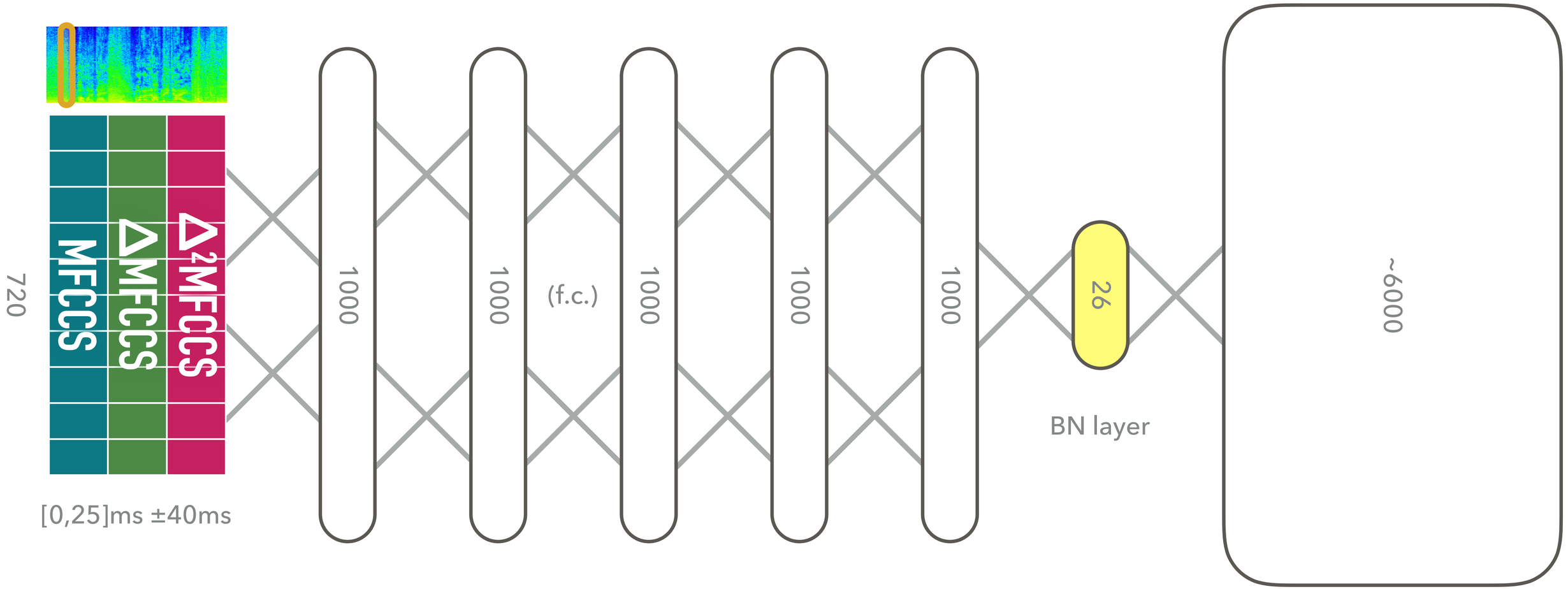
Moving forward: DNN-based ASR

(work in progress)

- DNNs have proved very effective in visual domain.
- Hidden-layer representations provide “bottom-up” features which are used to disambiguate speech.

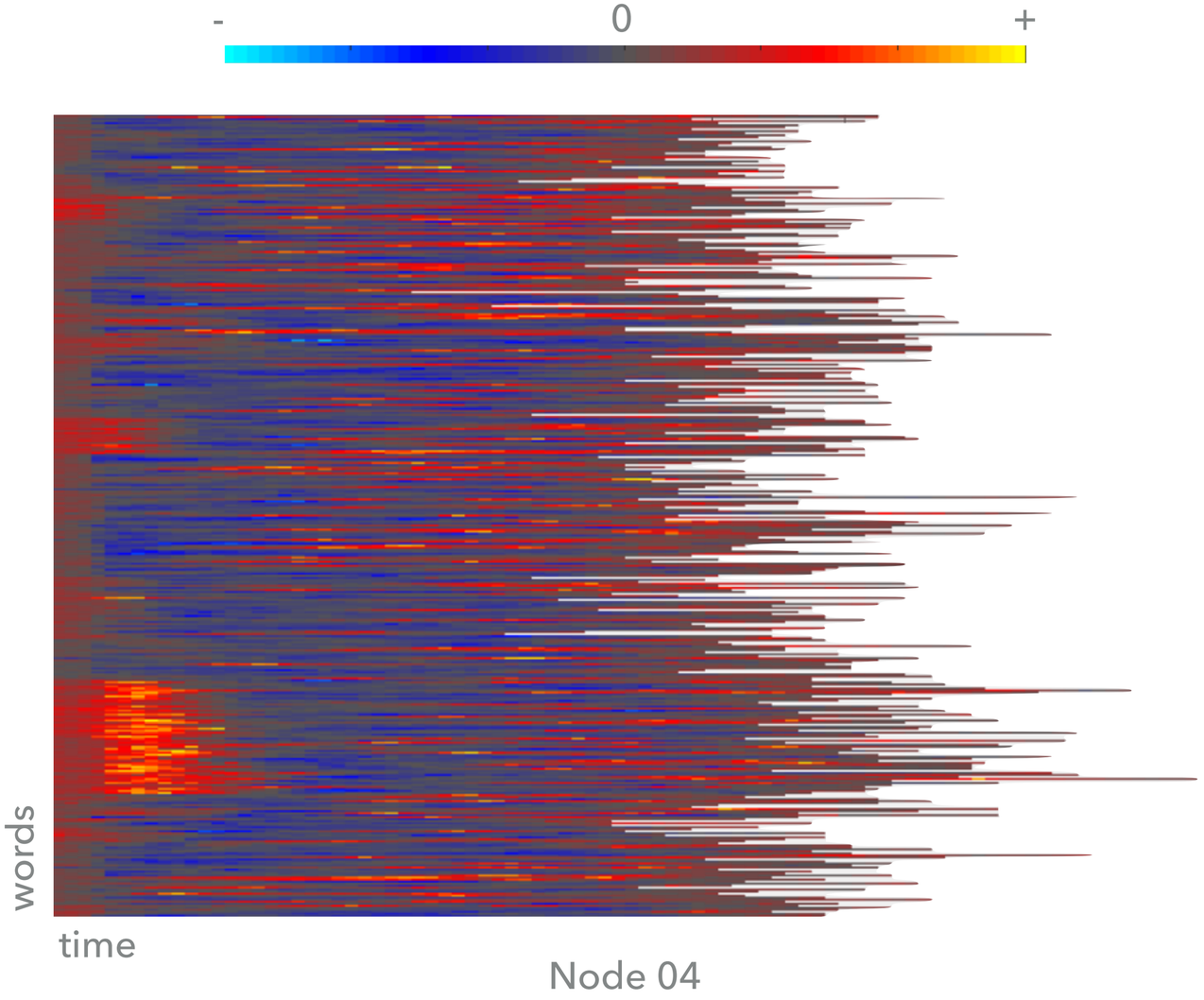
HTK: DNN-HMM

Zhang & Woodland (2015)
Submission to InterSpeech

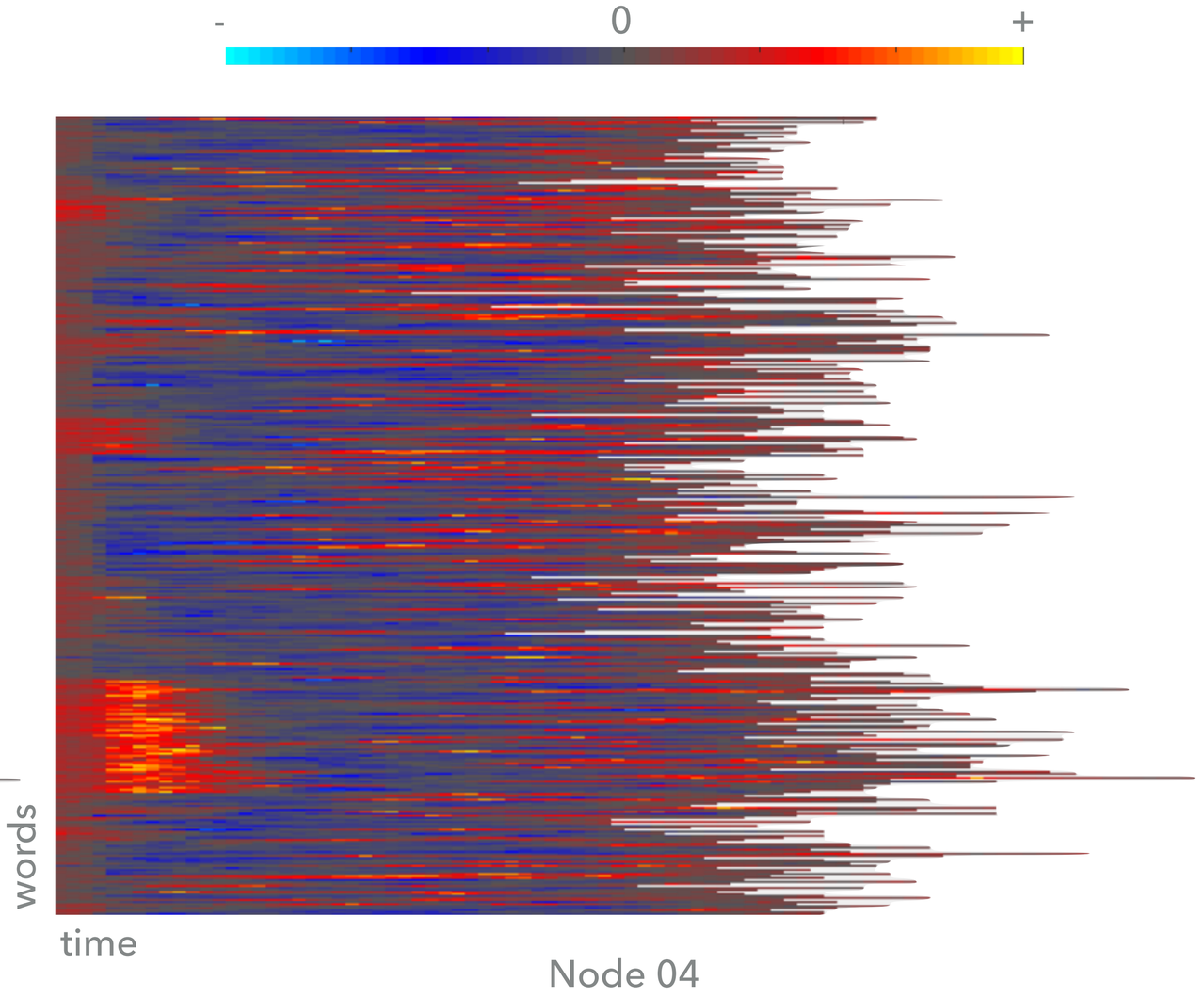


Individual node responses

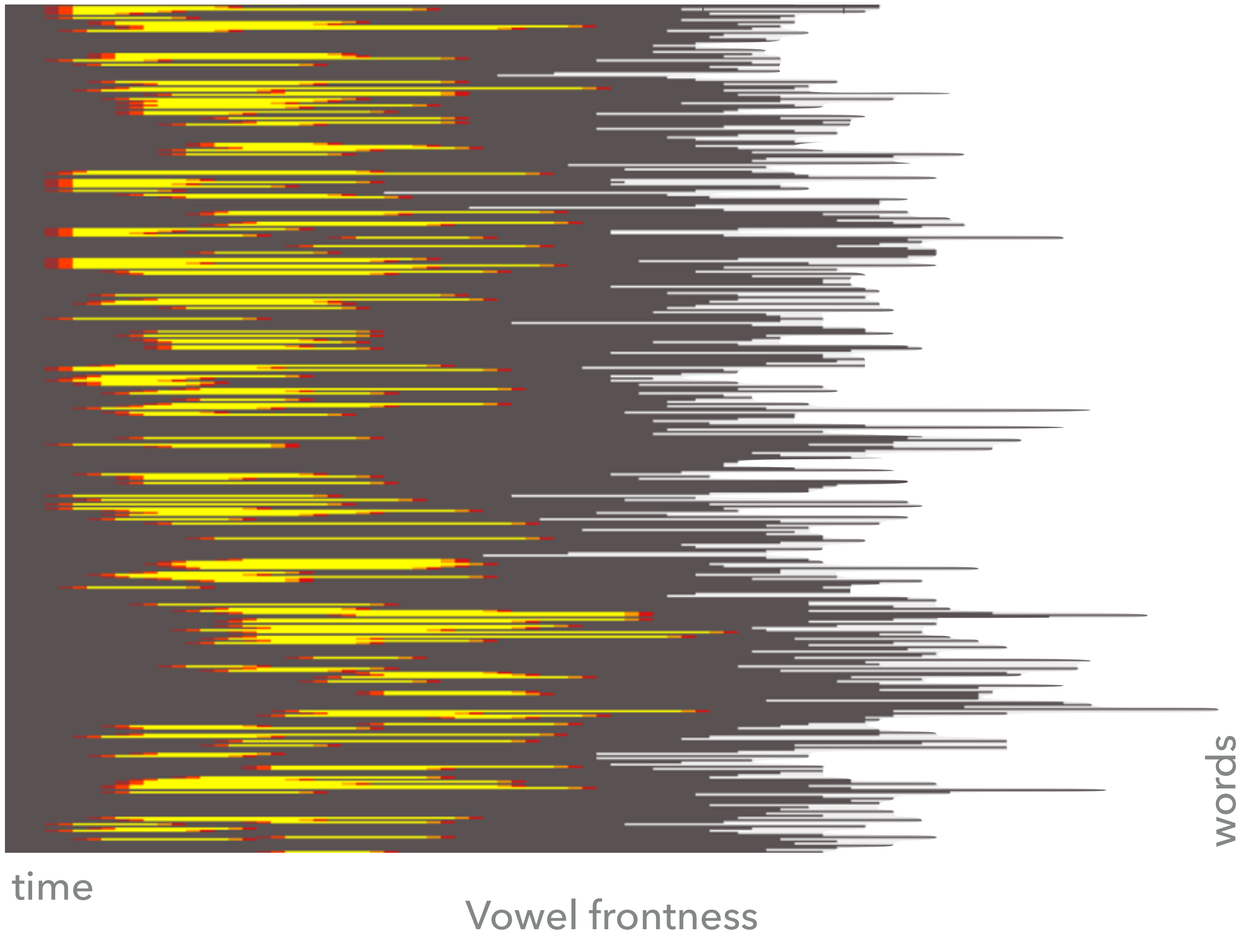
- BN architecture provides a low-dimensional feature space sufficient to accurately determine 6000+ phonetic labels.
- Dynamic inputs elicit dynamic BN responses.
- Can we investigate this BN representation space, and compare it to brain representations?

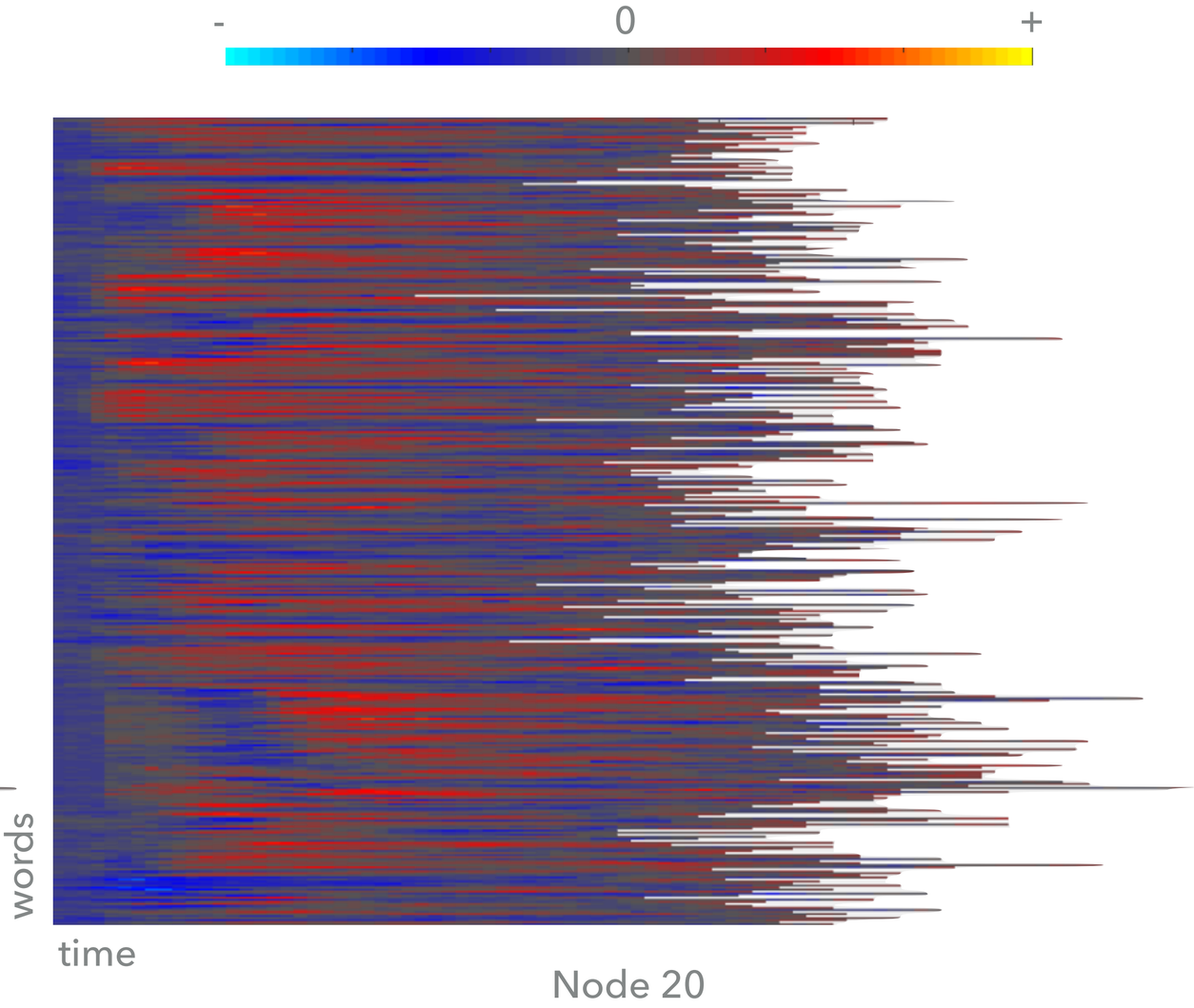


Nodes track phonetic features?

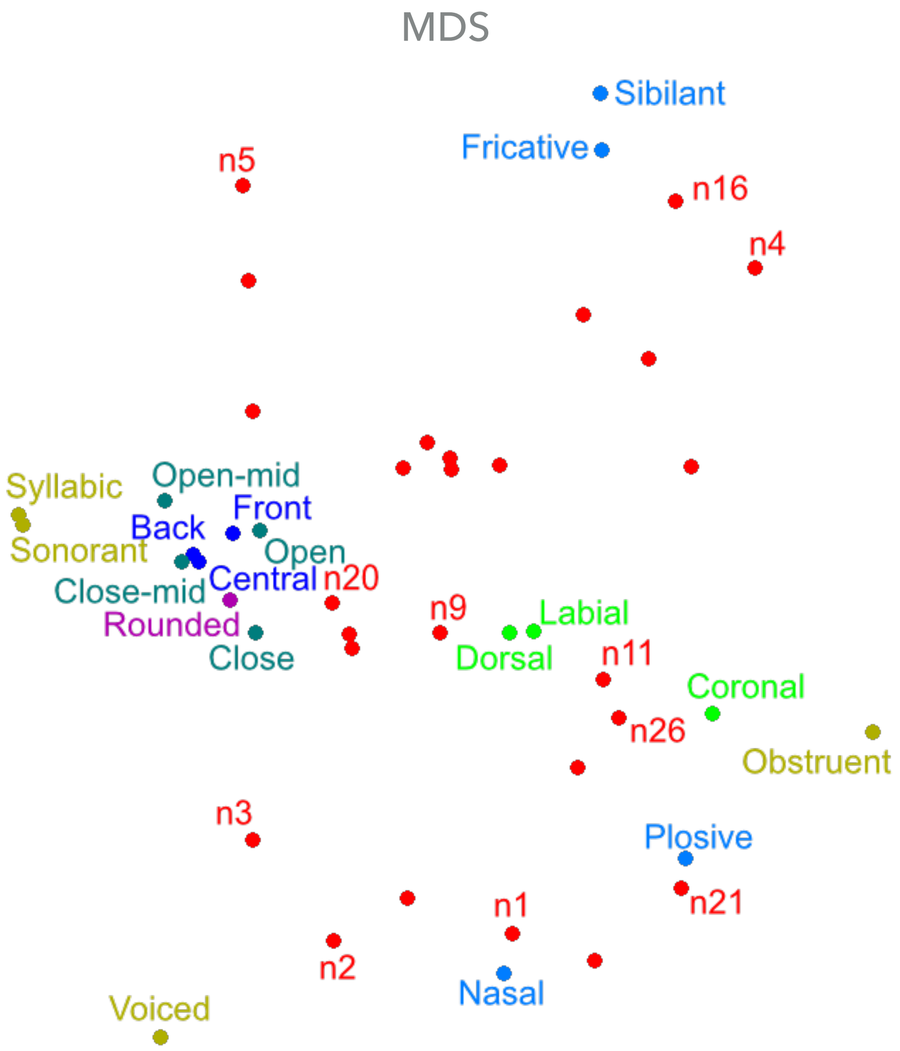


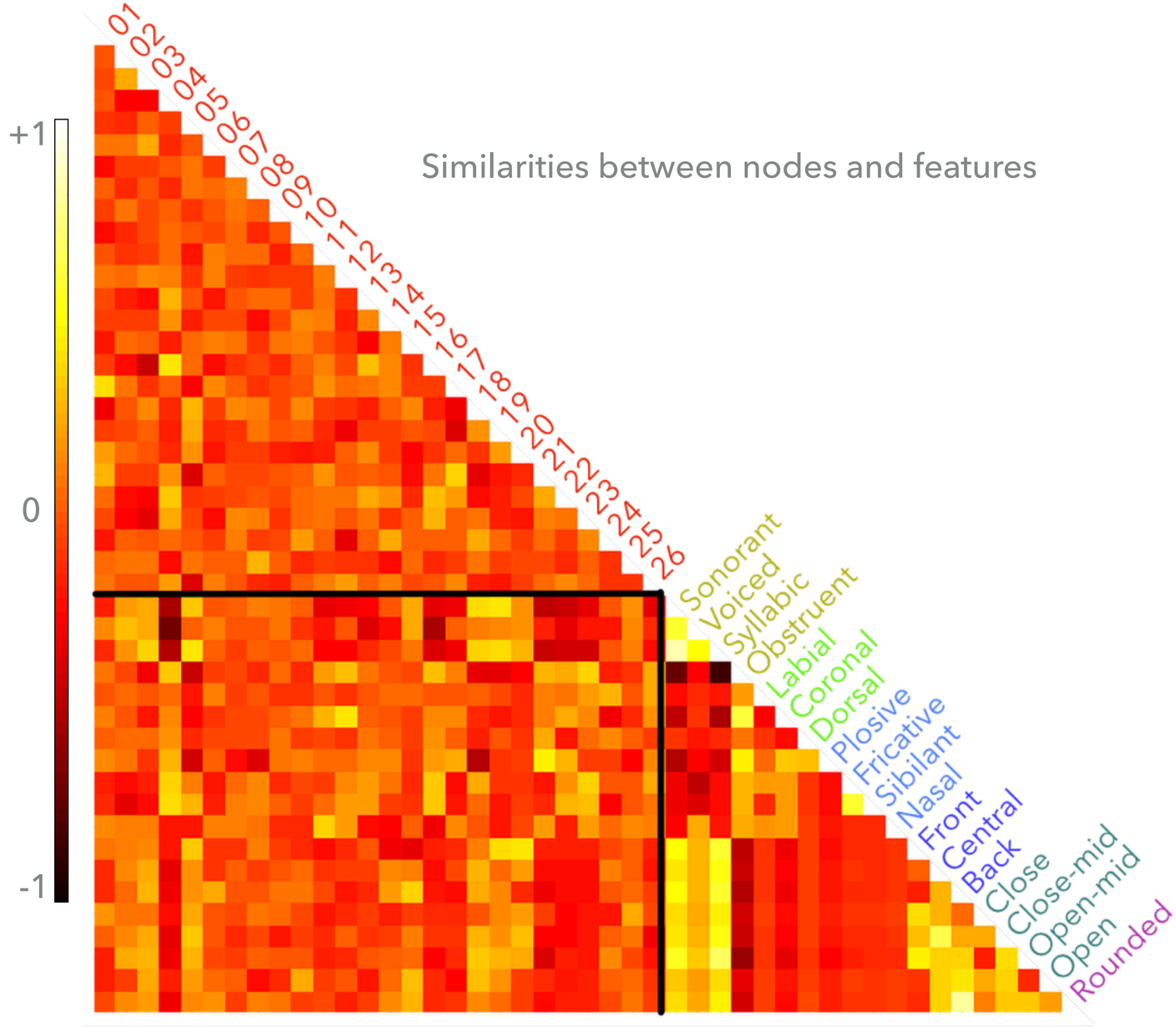
Nodes track phonetic features?





BN–feature similarity





Summary

- We found evidence of regions of articulatory feature representation in human auditory cortex.
- We modelled speech-recognition-relevant features using machine systems which perform the task well.
- RSA allows comparison of brain states and machine states at the level of representations.
- EMEG records rich brain response data over time, non-invasively.
- The processes of sound-to-meaning mapping are still poorly understood.


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