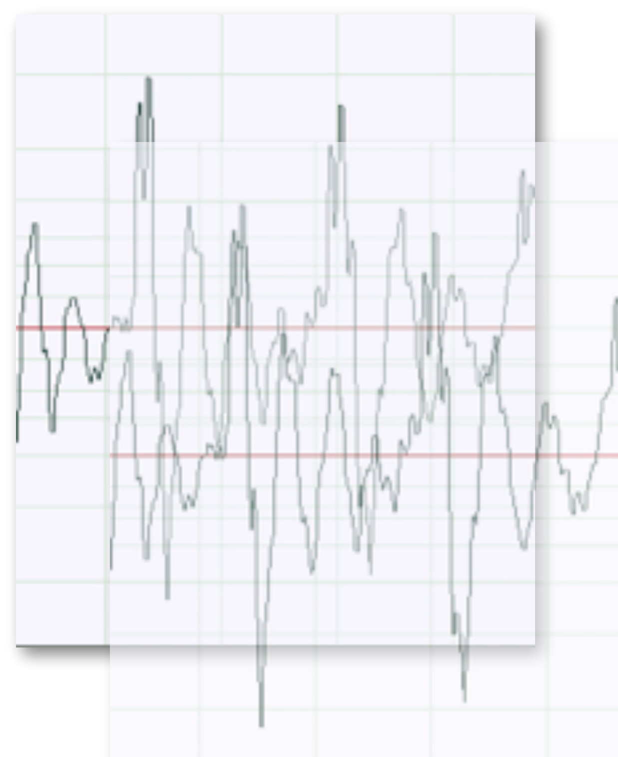
Reverse-engineering the cortical processing of speech

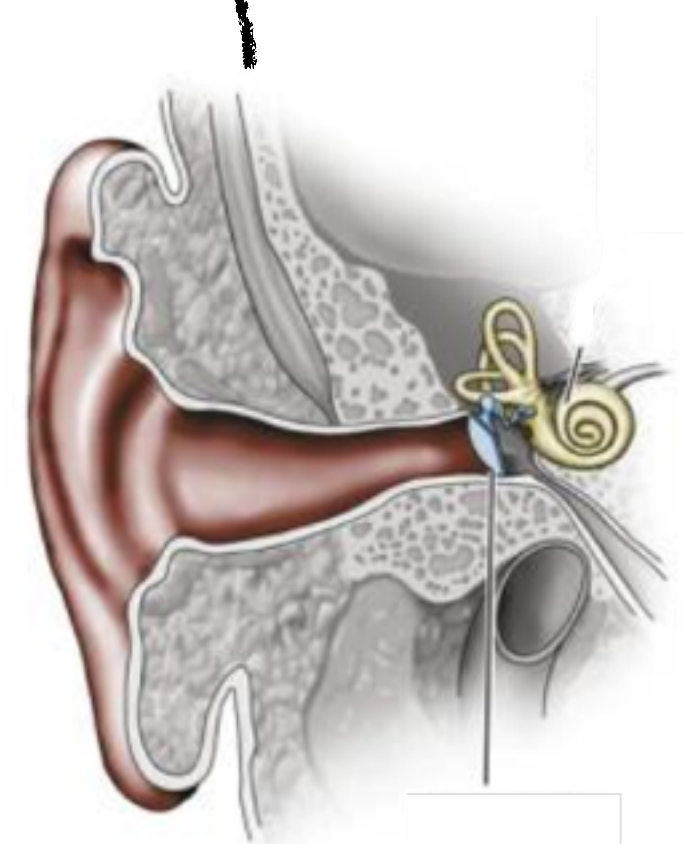
Nima Mesgarani
Electrical Engineering
Columbia University

Anatomy of speech communication

Production

Perception



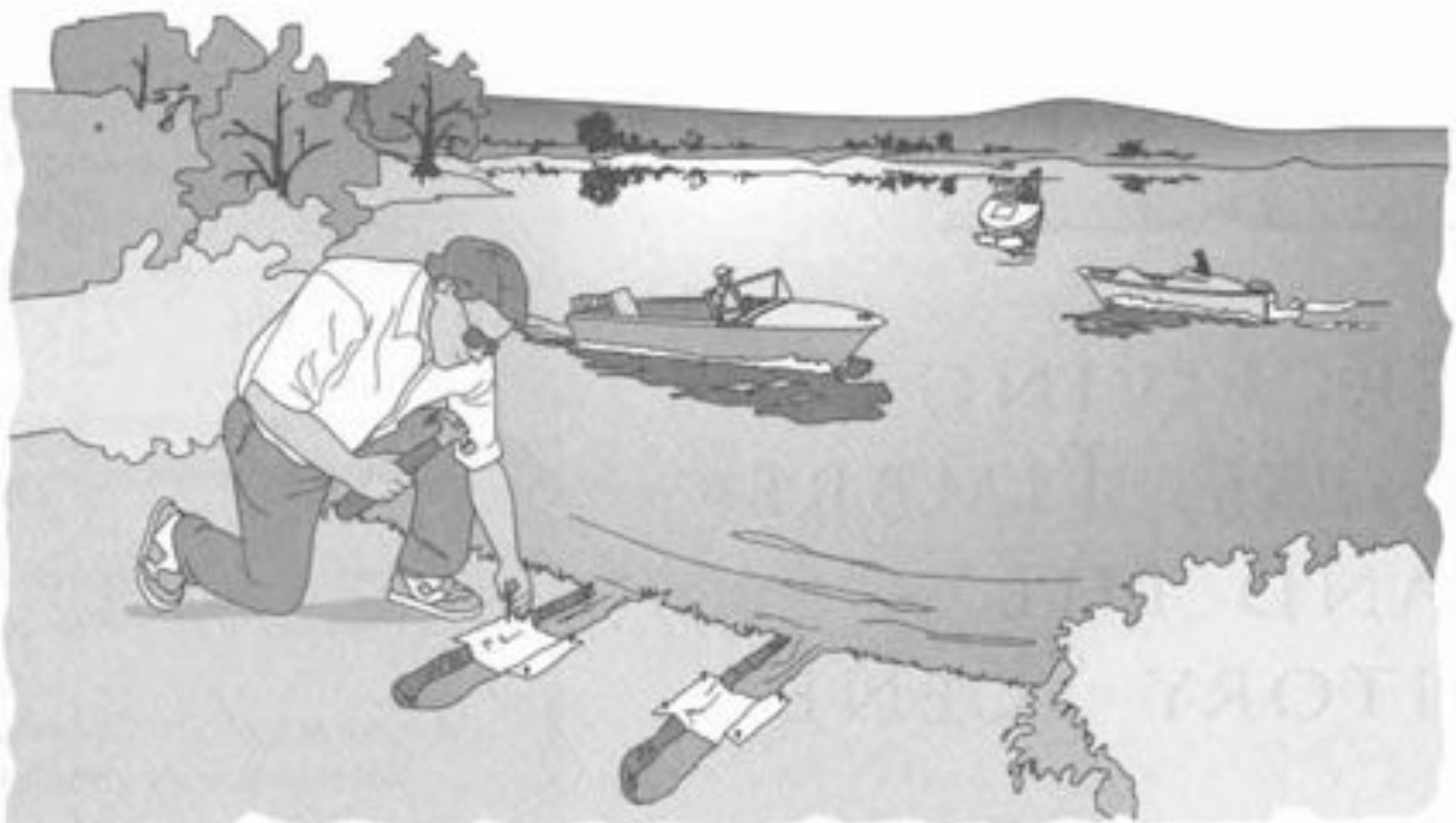


Ear drum

Cocktail party problem, Cherry, (1953)

Challenging problem

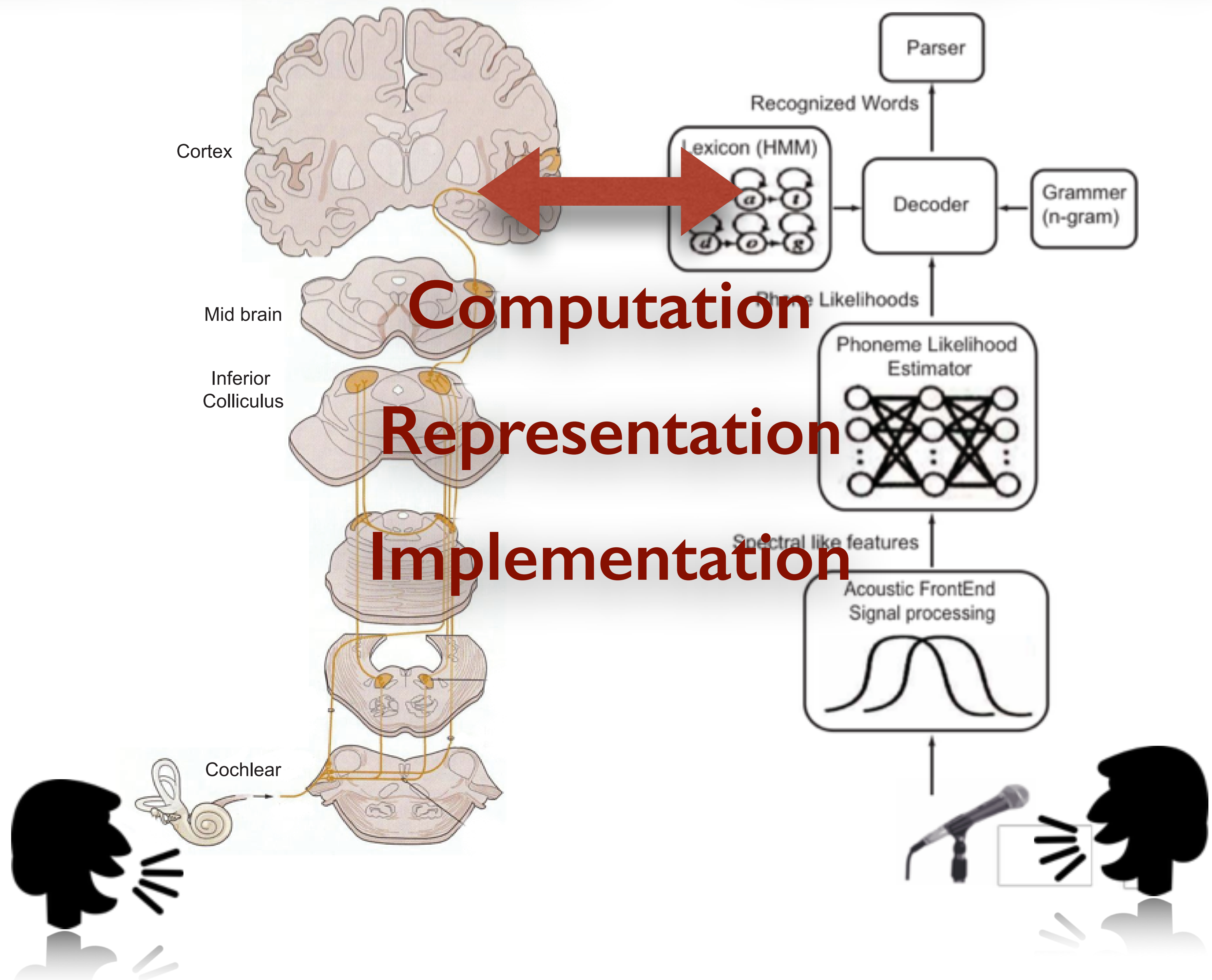
Segregating sound mixtures into separate perceptual acoustic events (sources) (Bregman'90)



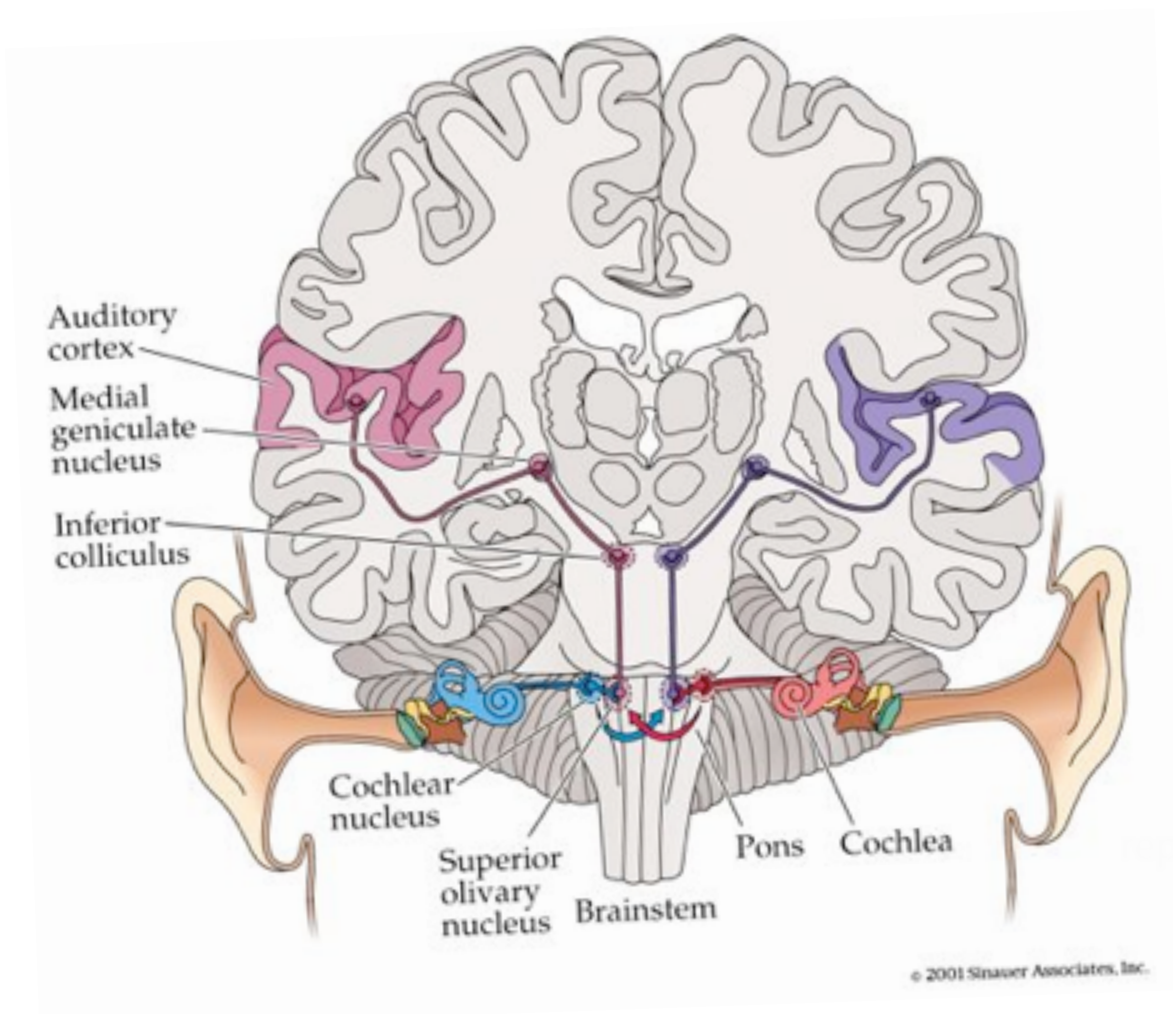
Creating a model of speech communication

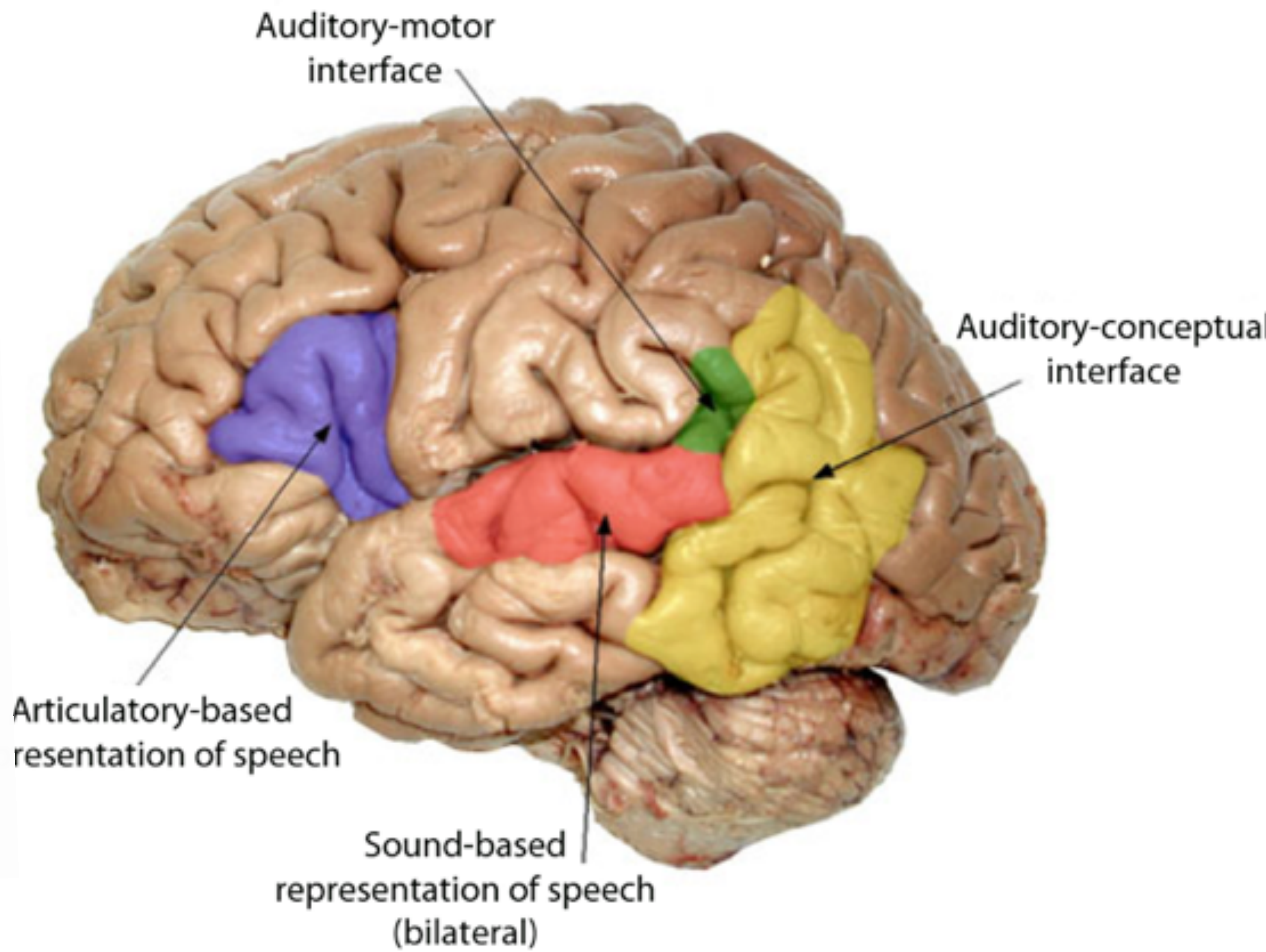
Understanding the brain, speech disorders, prosthesis

Closing the gap between artificial and biological computing



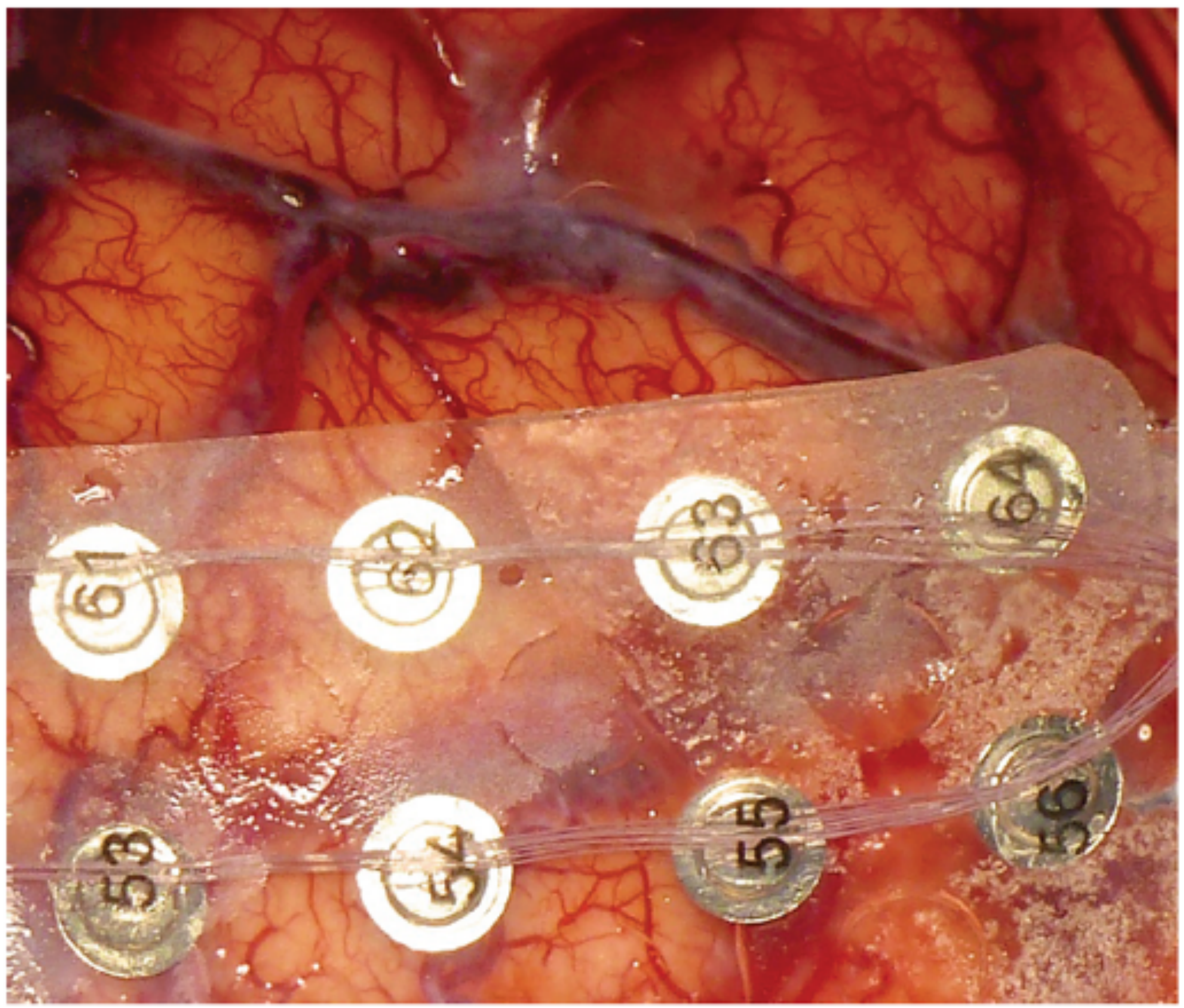
Speech processing in the brain



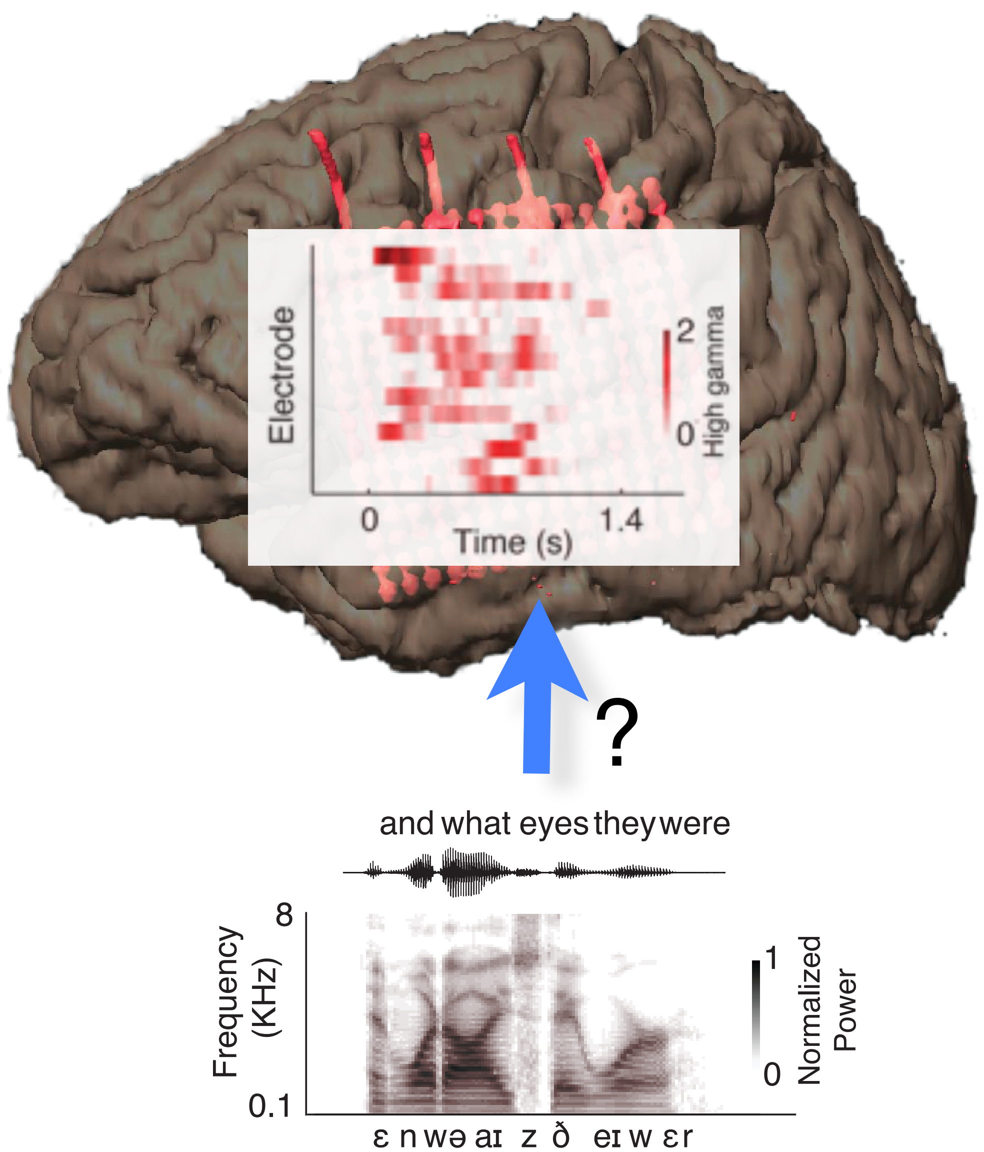


ElectroCorticoGraphy (ECoG)

- Implanted chronic grids for localization of epileptogenic foci usually 7-10 days.



Making sense of the cortical representation of speech

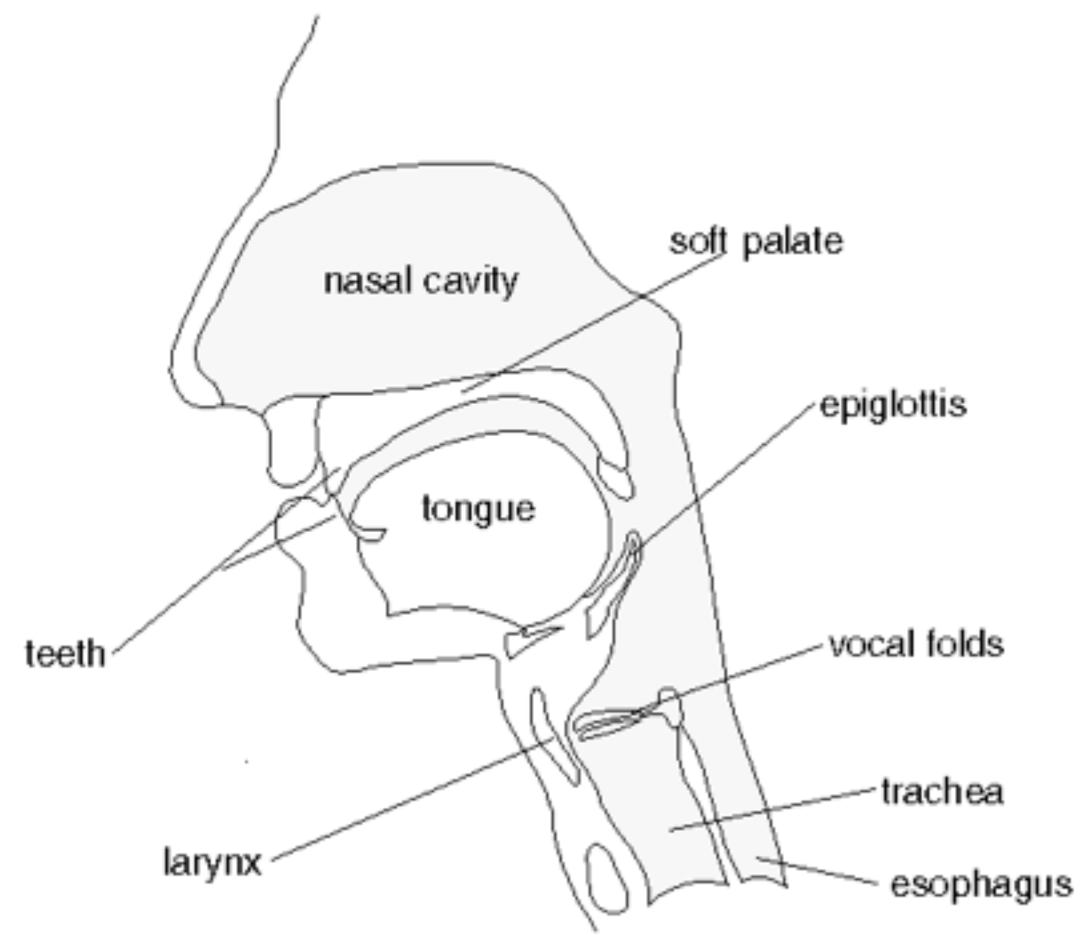


500 unique sentences from 400 speakers

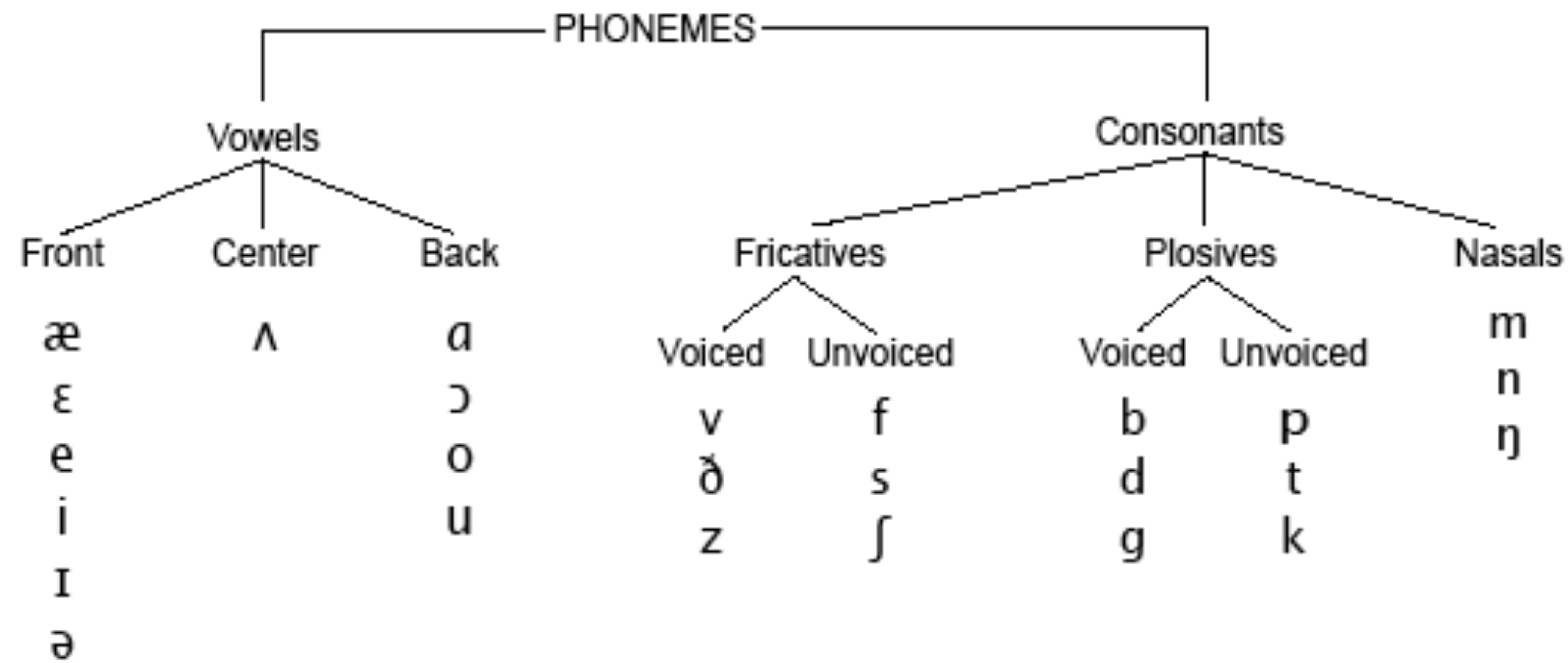
- Are different auditory sites selective to specific speech sounds?
- What features organize the neural responses?
- What natural variabilities are encoded?

Using phonemes to segment speech

- Smallest contrastive linguistic unit that can change the meaning
  - /b/ in /bad/
    /d/ in /dad/
- Limited inventory in each language

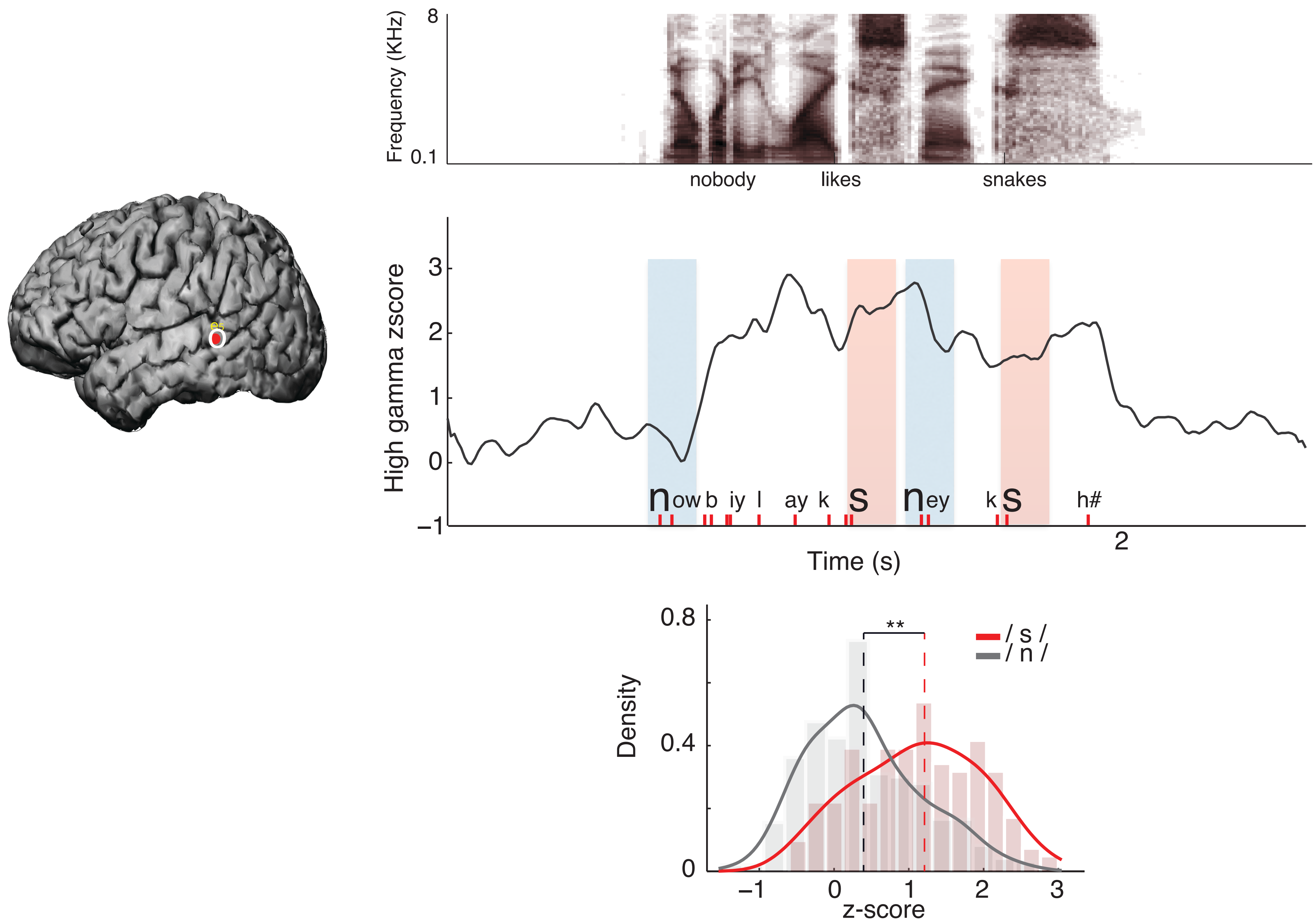


Phonetic categories

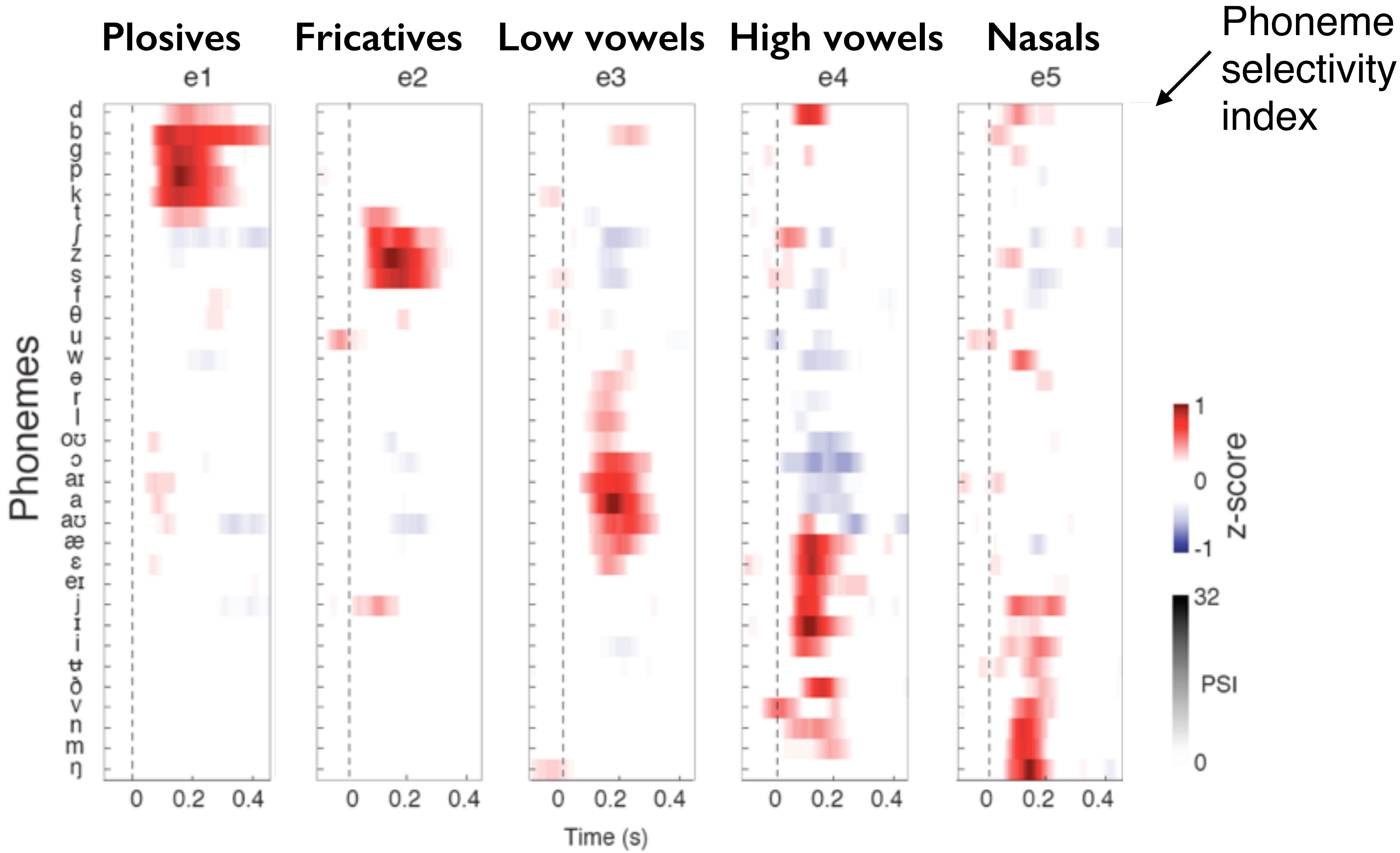


Distinctive features, Chomsky, Halle, Stevens

Specificity of neural responses



Examples of average phoneme responses in STG



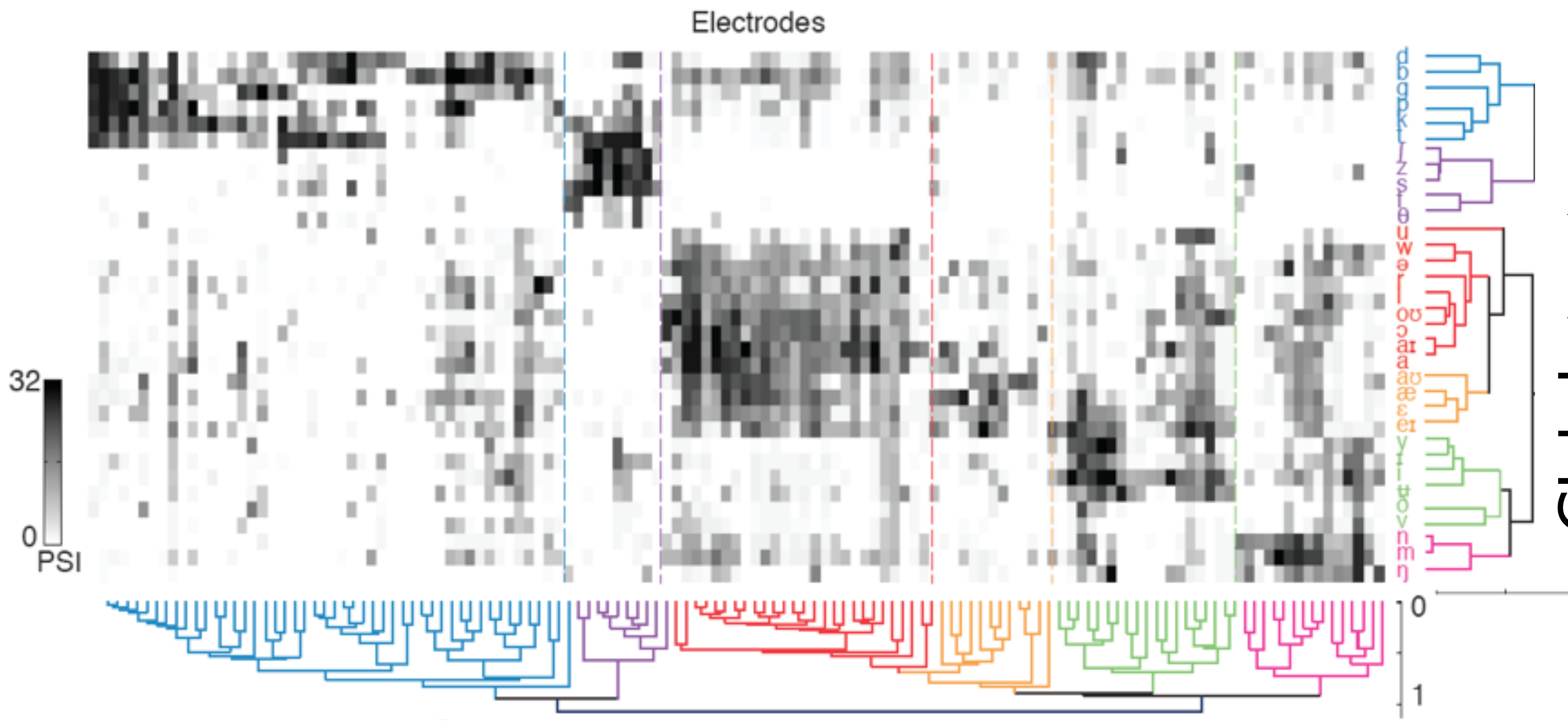
Diversity of responses: Strong preference at various STG sites to specific phoneme **groups** with shared attributes

Mesgarani et. al, 2014, *Science*

Selectivity pattern across all STG sites

What ‘types’ of selectivity patterns at local and population level?

Clustering the PSI vectors



Global structures (population)

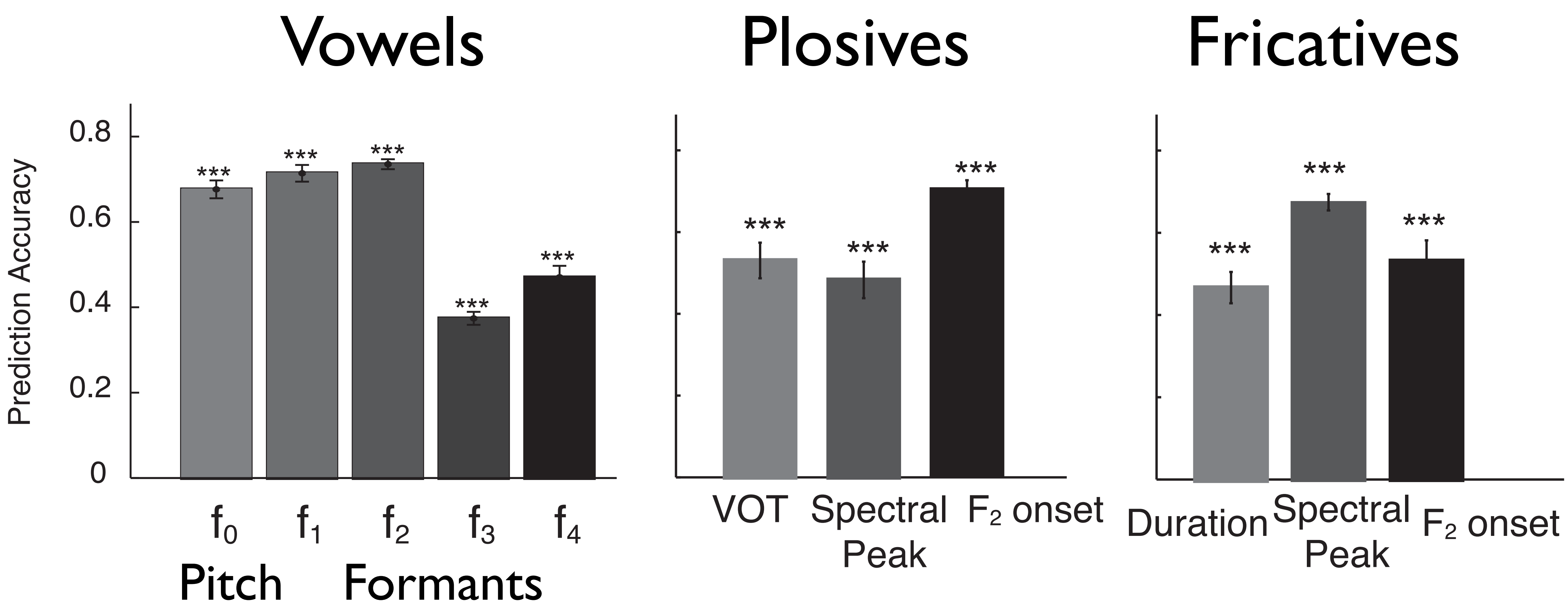
Local structures (single electrode)

Place

Manner

Mesgarani et. al, 2014, *Science*

Prediction accuracies of acoustic parameters of phones from population

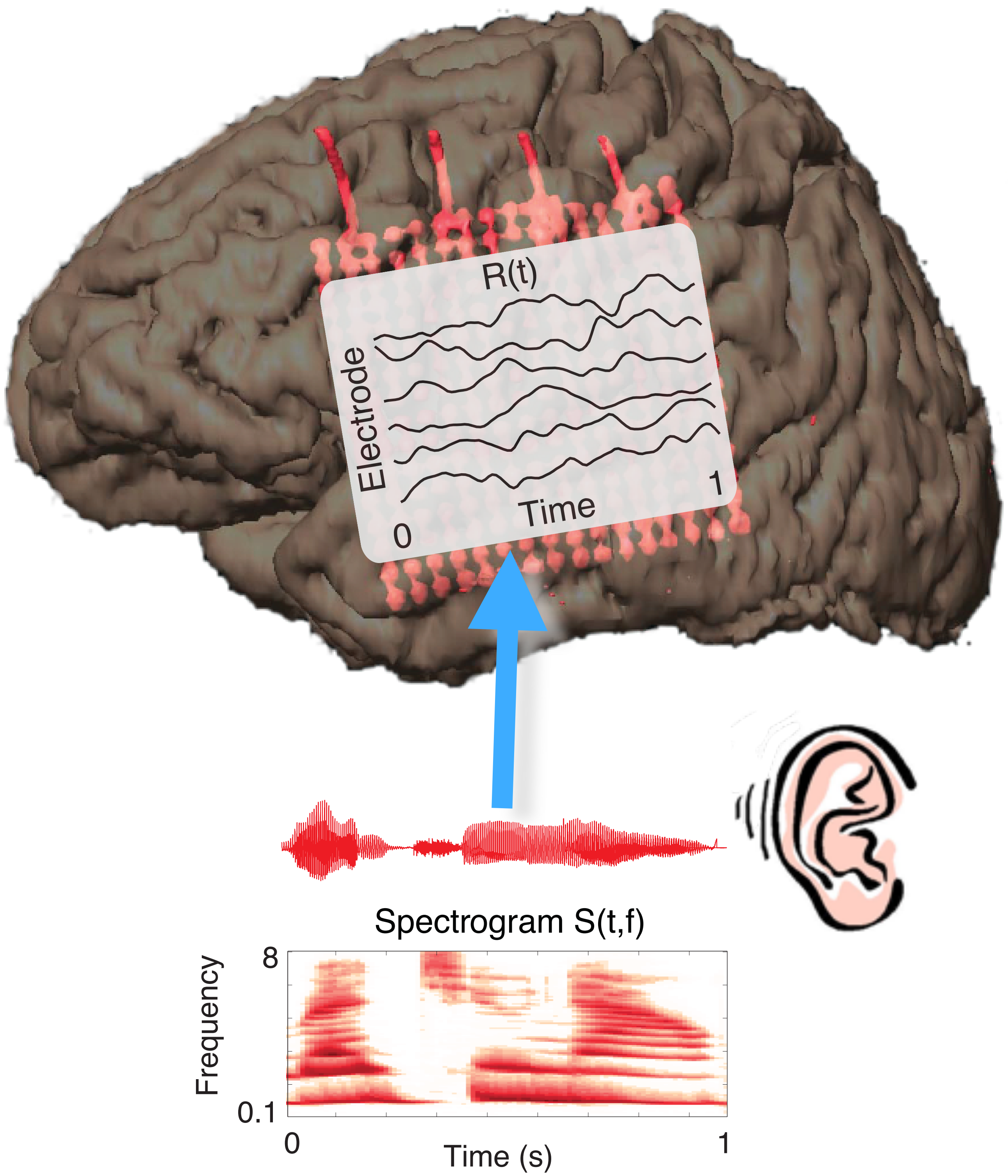


The natural variability of phones is encoded in STG responses

Representation of speech in STG

- Single electrode selectivity to phonetic feature categories (e.g. place and manner)
- Accurate encoding of natural variabilities of phones
- Evidence for nonlinear encoding of Voice-Onset-Time, and joint encoding of formant frequencies

Inverse model: From neural response to sound

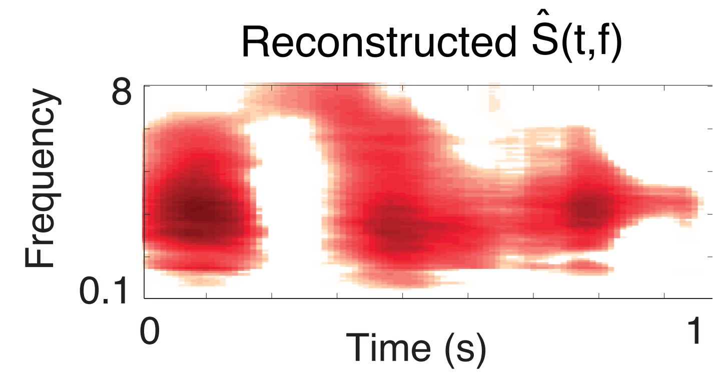


$$R(t) \xrightarrow{G(t,f)} S(t,f)$$

$$S(t,f) = \sum_{n} \sum_{\tau} G_n(t-\tau, f)\, r_n(\tau)$$

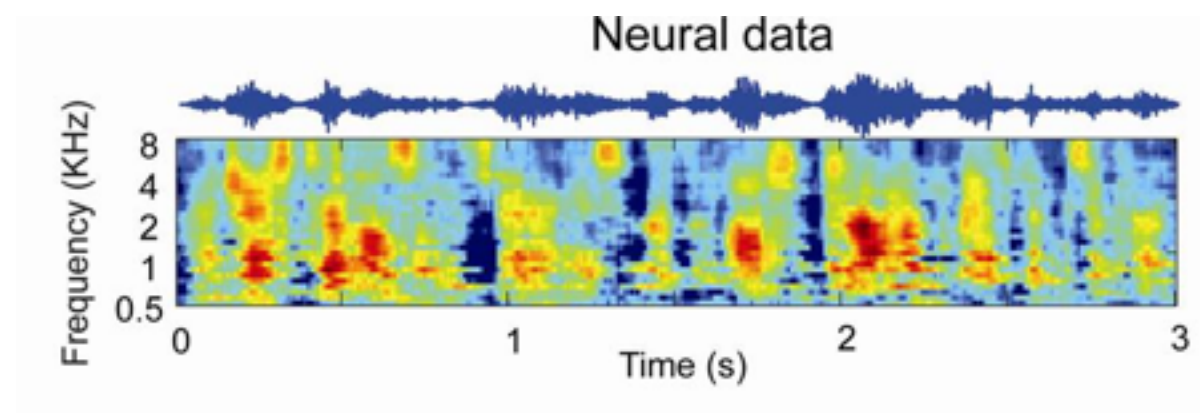
$$e = \sum_{f} \sum_{t} (\widehat{S(t,f)} - S(t,f))^2$$

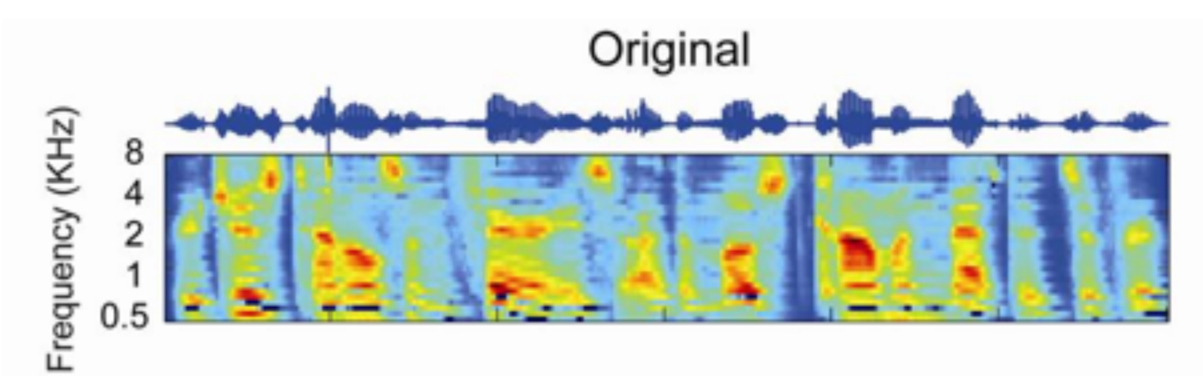
$$G = C_{rr}^{-1} C_{rs}$$



Mesgarani et. al J. Neurophysiology 2009

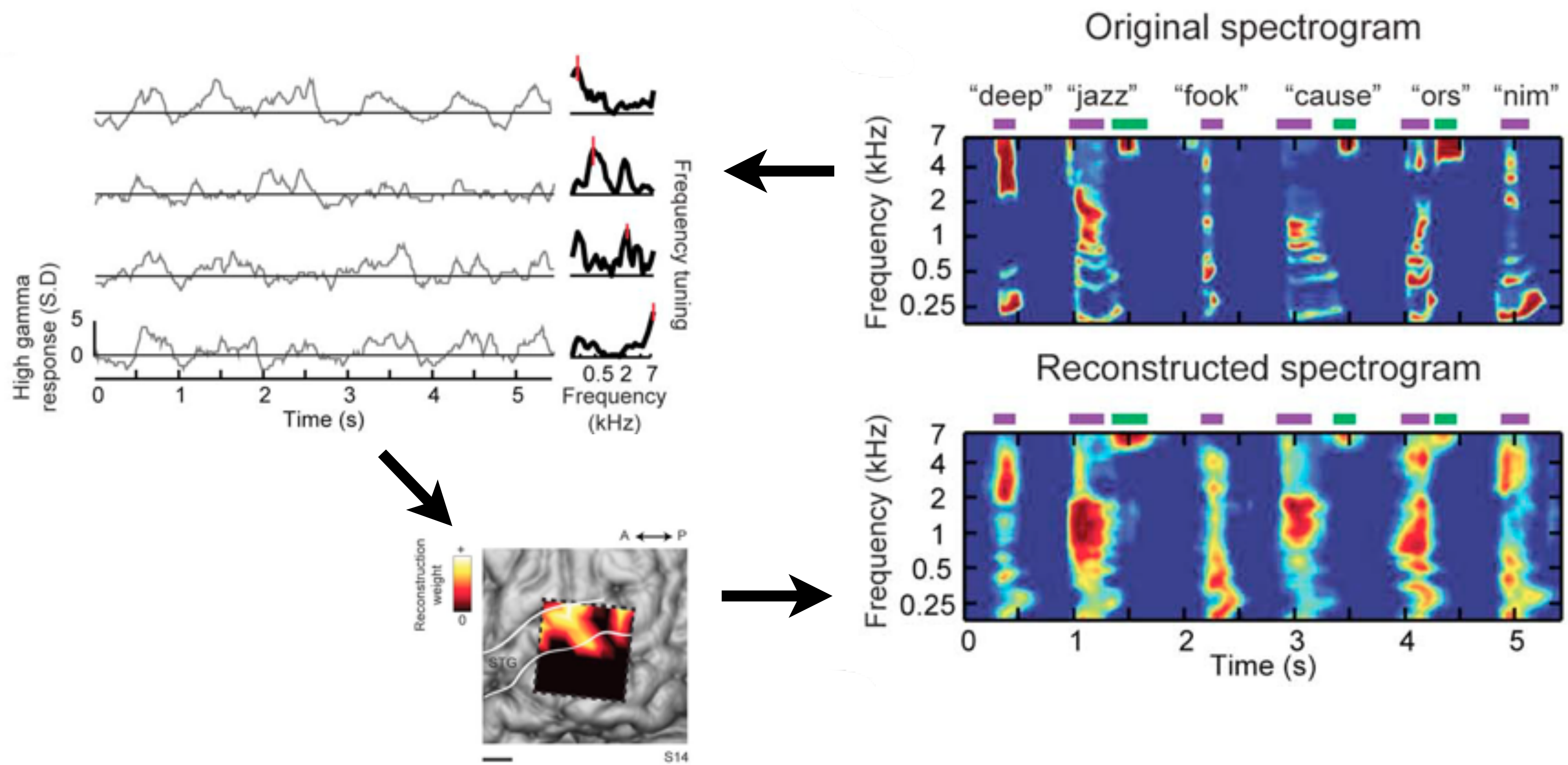
Reconstruction from 200 single units in Ferret A1





Mesgarani et.al J. Neurophysiology 2009

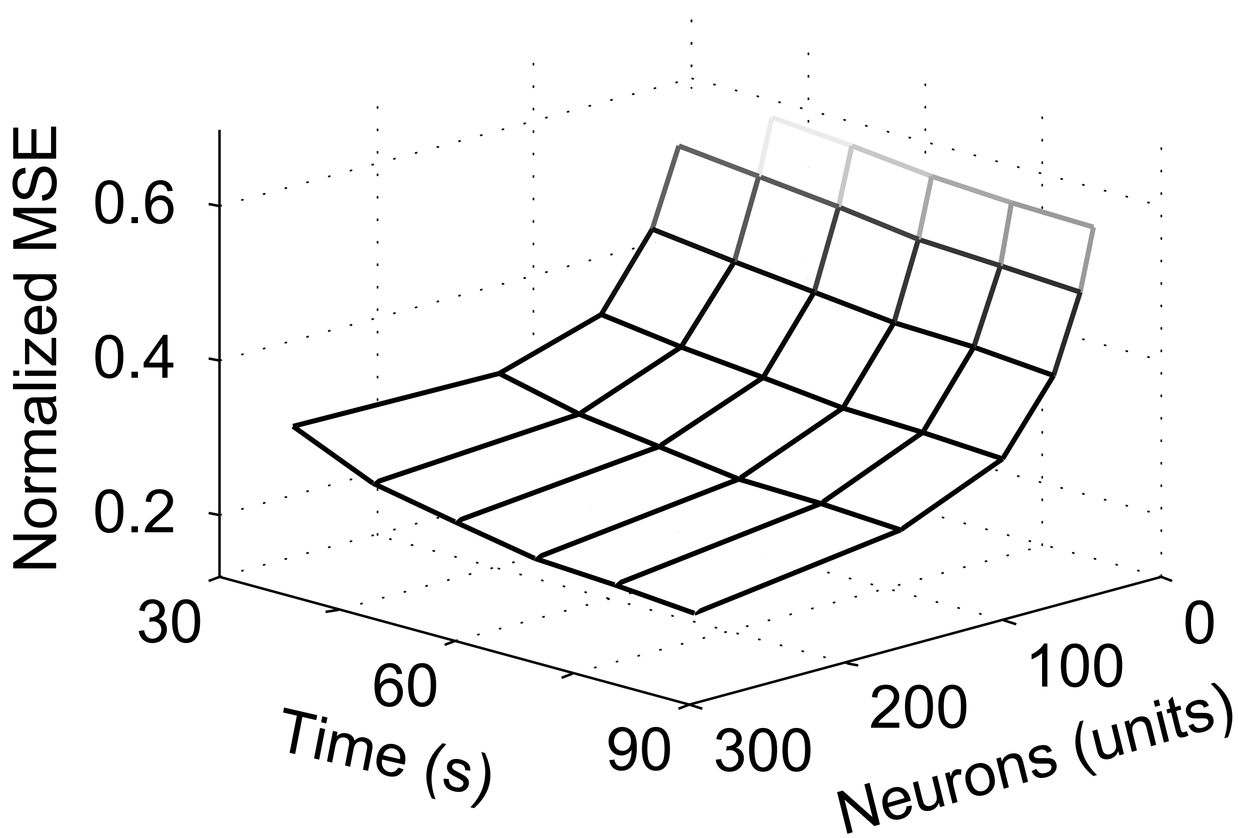
Reconstruction from human brain



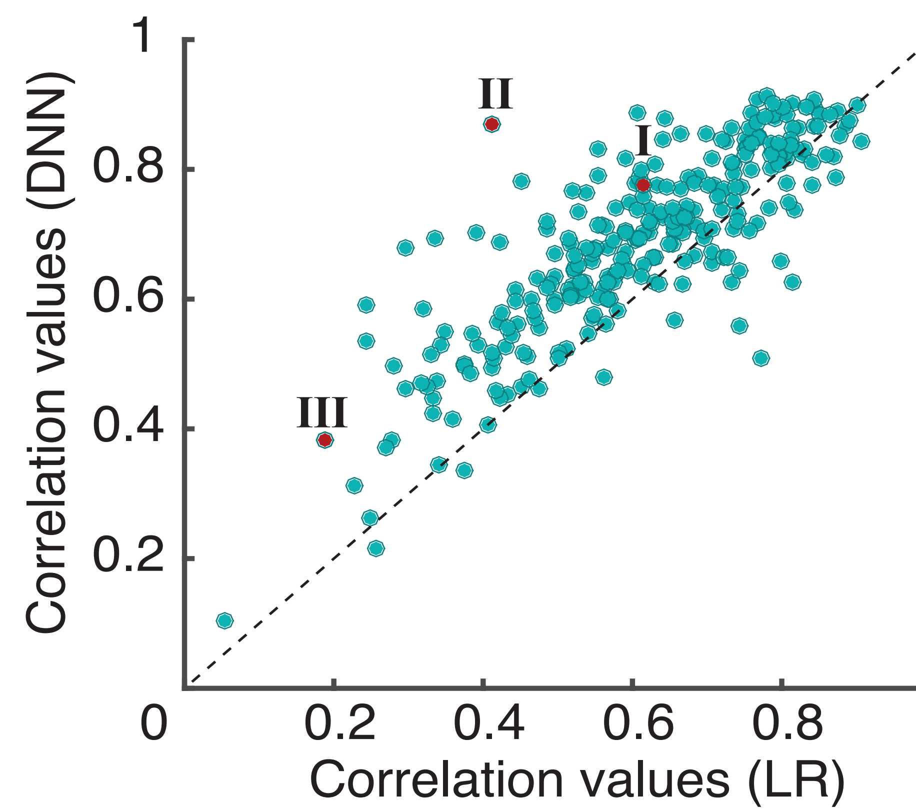
Pasley et. al. PLoS Biology, 2012

Improving reconstruction accuracy

More training data
and more neurons

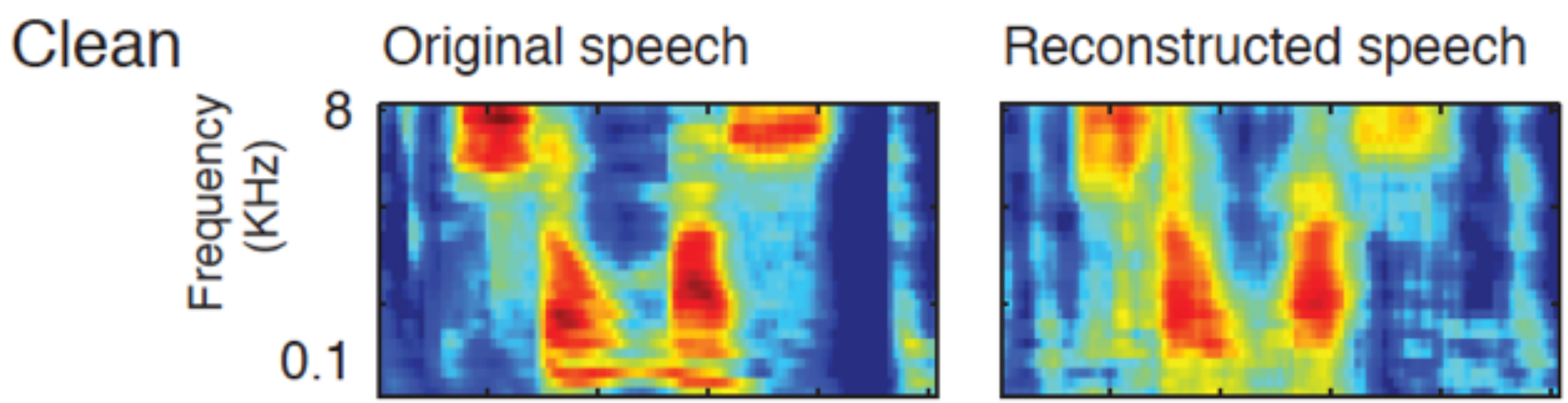


More advanced models
(DNN)



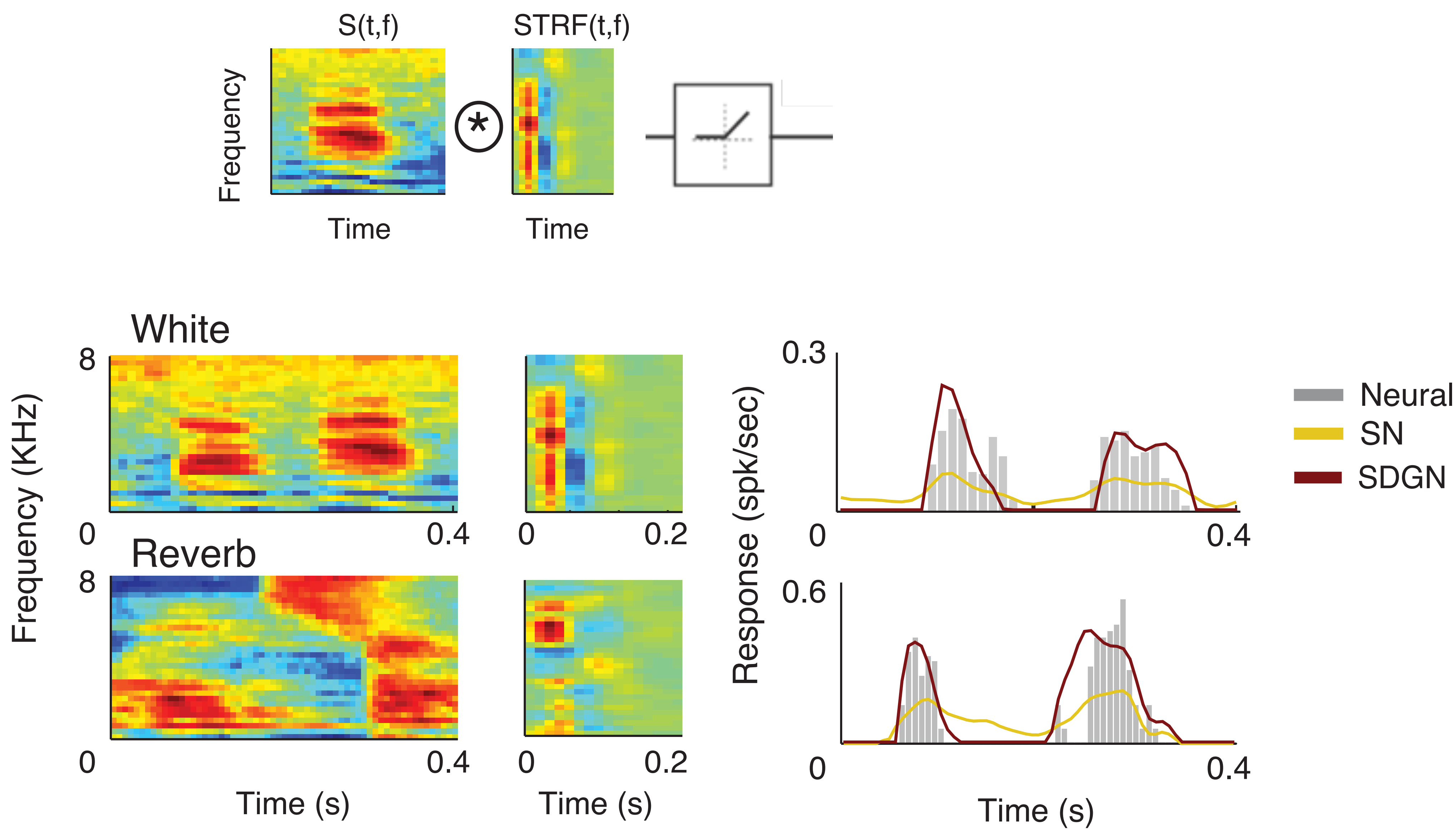
Yang et. al, Interspeech 2015

Reconstructing noisy speech from auditory cortex



Mesgarani et. al. *PNAS* 2014

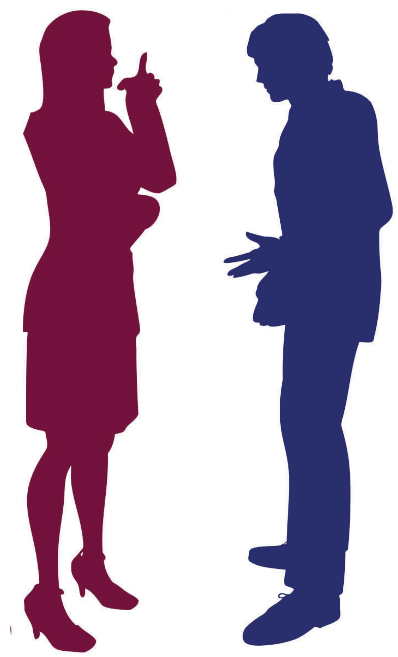
A dynamic model of auditory cortical neurons

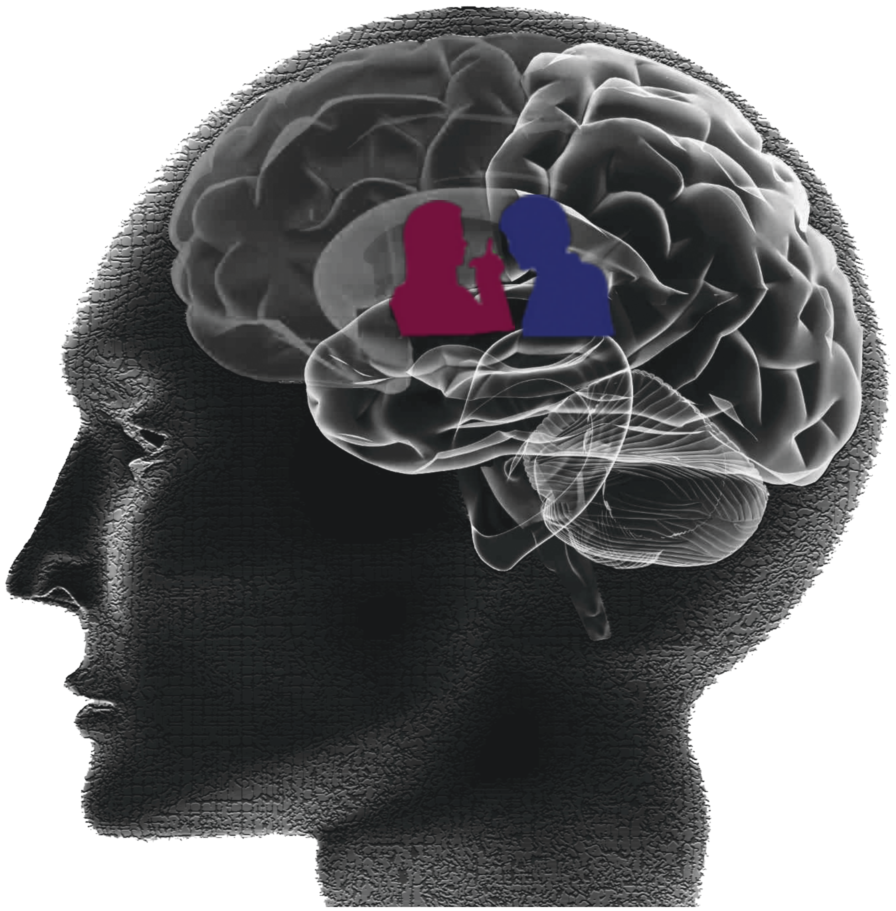


Static model CANNOT account for noise robustness of neural responses

Mesgarani et. al., *PNAS* 2014

Attentional modulation of the cortical representation



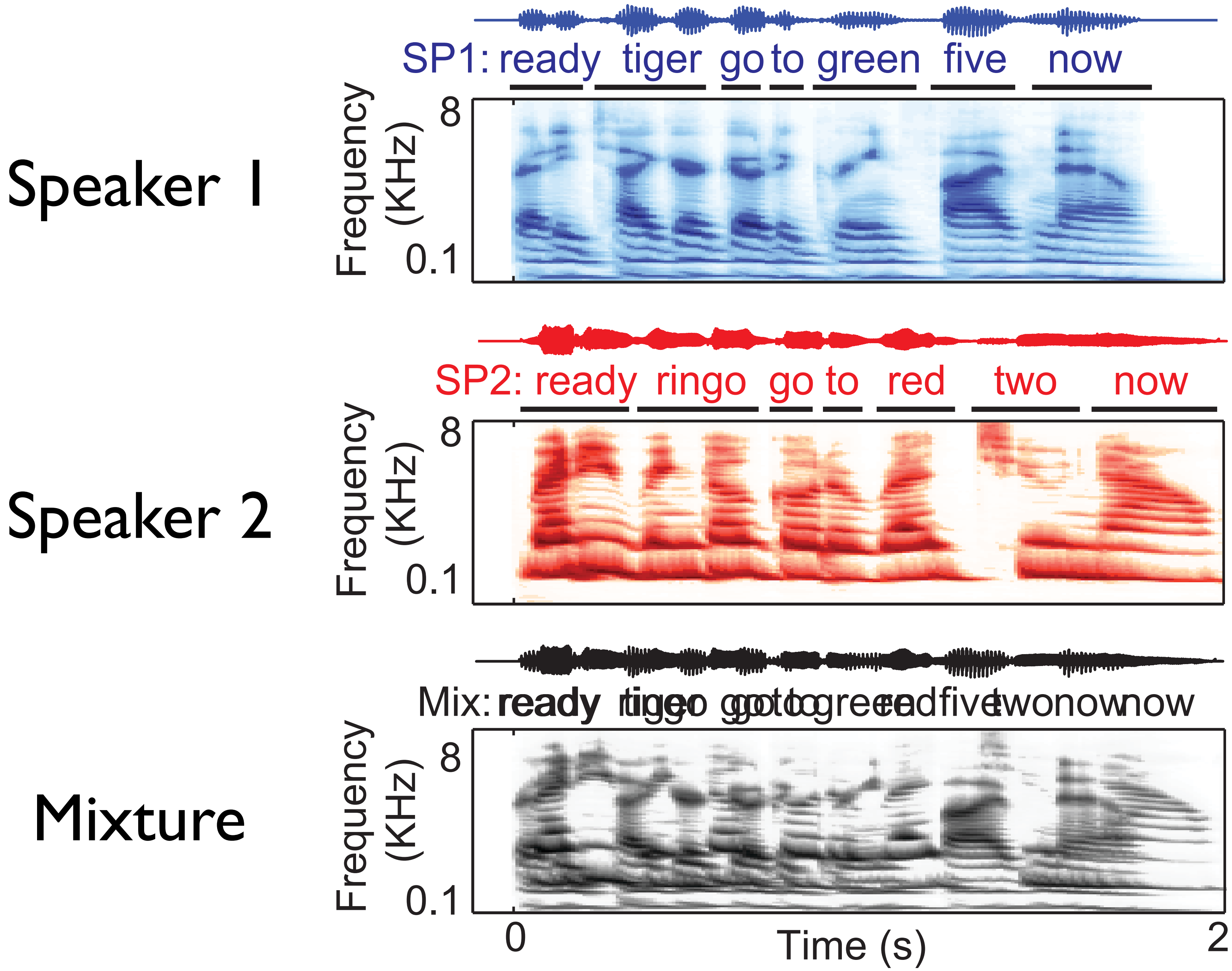


- What is the representation of attended speaker?
- Neural correlate of perceptual failures

Cocktail party problem, Cherry, (1953)

Experiment design

“Ready [Call Sign] go to [Color] [Number] now”



Experimental setup

Visual:
Target Call Sign

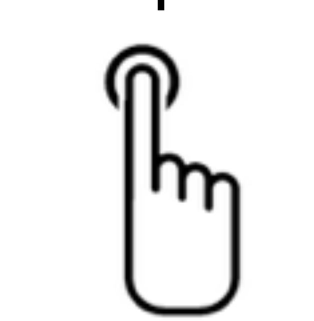
Ringo

Listen: Two simultaneous speakers

Target : “Ready Ringo go to [Red] [Two] now”

Distractor: “Ready Tiger go to [Green] [Five] now”

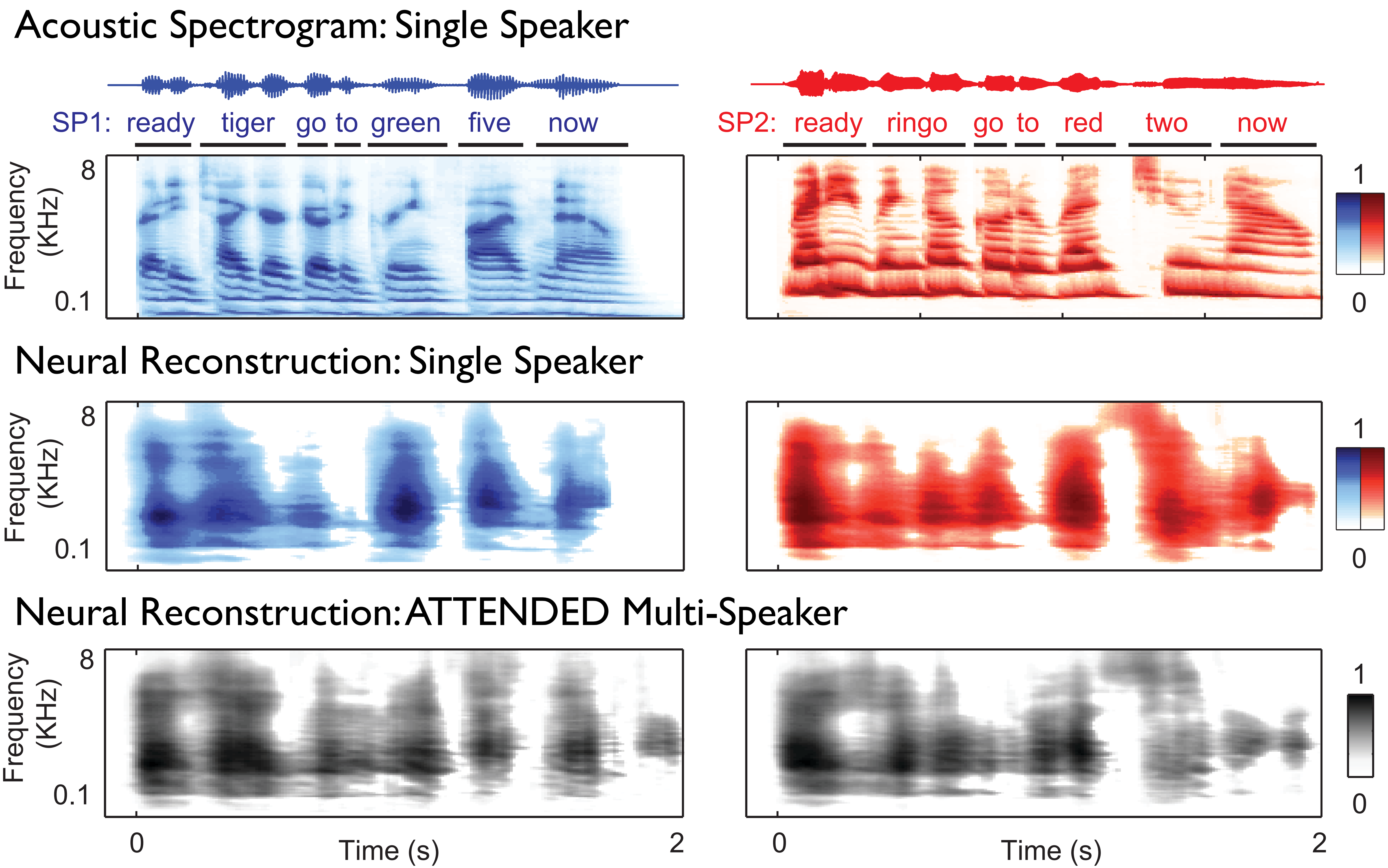
Response



Blue	**Two**
Red	Five
Green	Seven

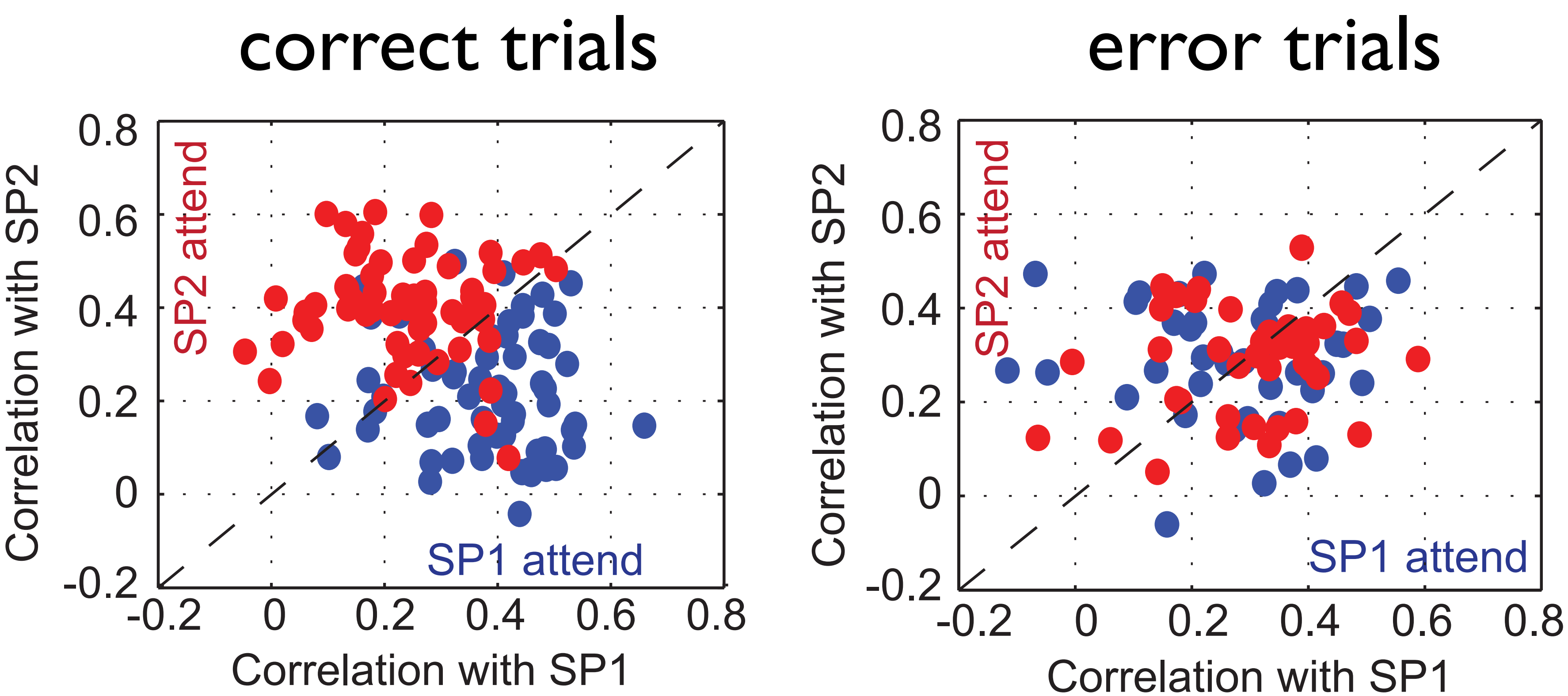
- Target speaker changes randomly from trial to trial
- Target call sign changes after each trial block

Attentional modulation of cortical representation



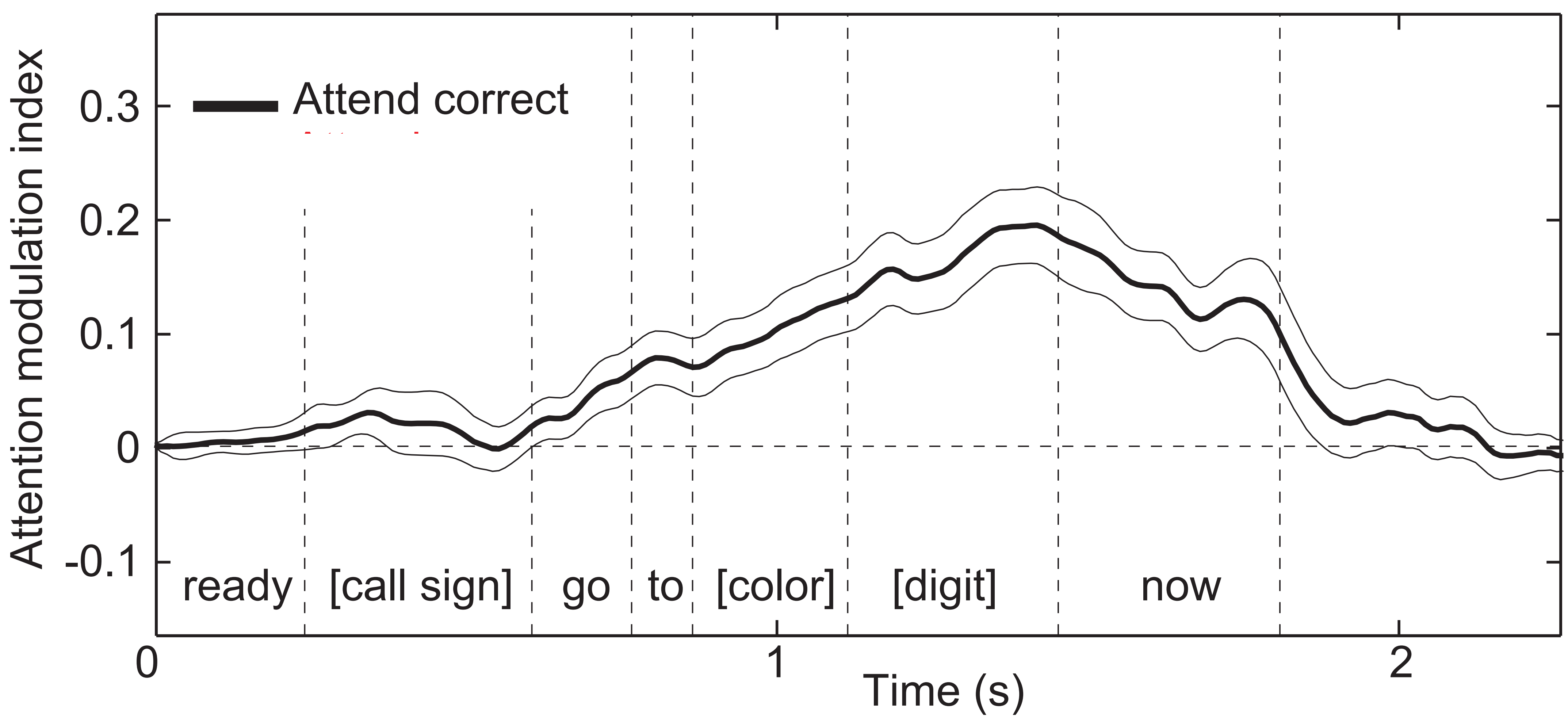
Mesgarani & Chang, (2012), *Nature*

Correlation with single speaker spectrograms



Mesgarani & Chang, (2012), *Nature*

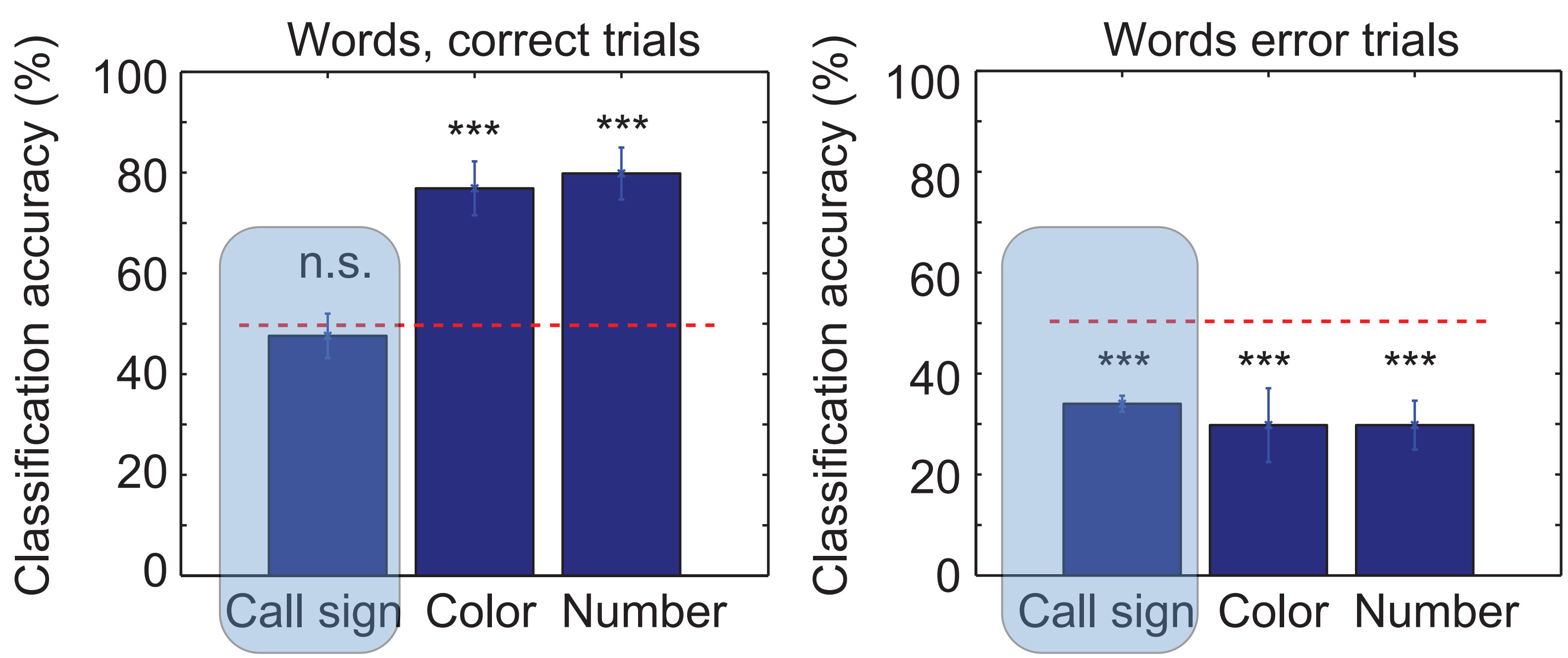
Time-course of attentional modulation



$$AMI = Corr\left(SP_1\ attend, SP_1\ alone\right) - Corr\left(SP_1\ attend, SP_2\ alone\right)$$
$$+Corr\left(SP_2\ attend, SP_2\ alone\right) - Corr\left(SP_2\ attend, SP_1\ alone\right)$$

Decoding words and identity of attended speaker using single speaker models

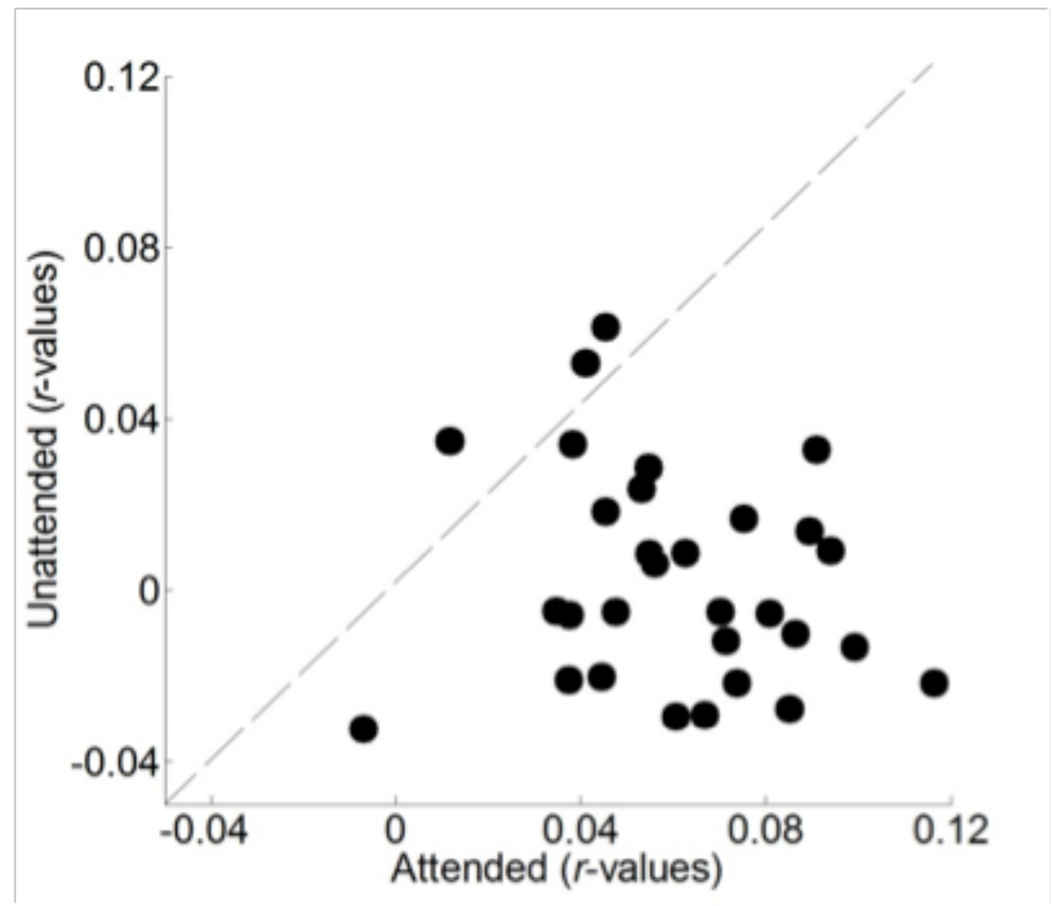
- Train linear, frame-based, classifier (RLS) on examples of single speakers responses and then decode the mixture speech



Online decoding of attention using single-trial EEG

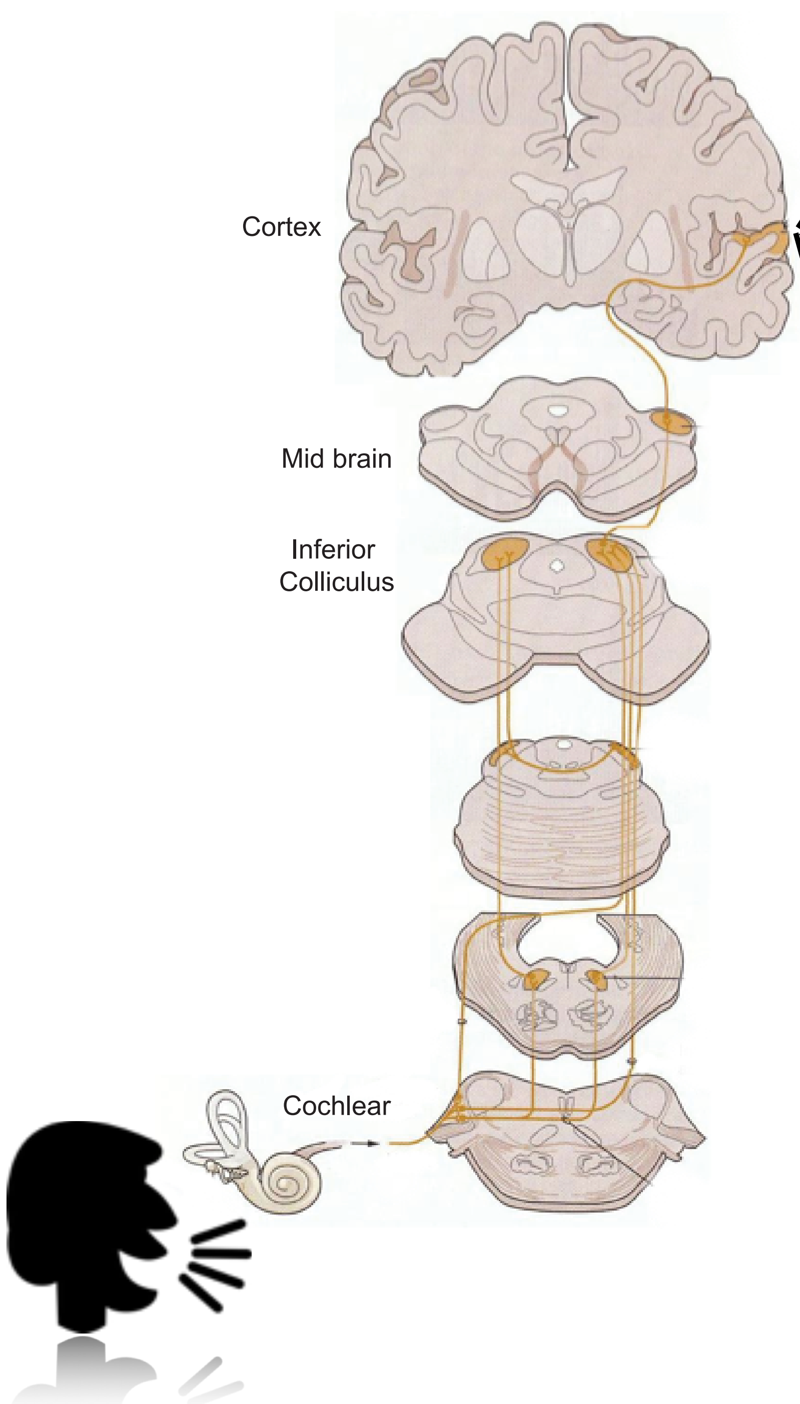
EEG recording setup

Similarity to attended speaker



Sulivan et. al., Cerebral Cortex, 2014

Cortical representation of speech



Selectivity to phonetic feature categories (e.g. place and manner)

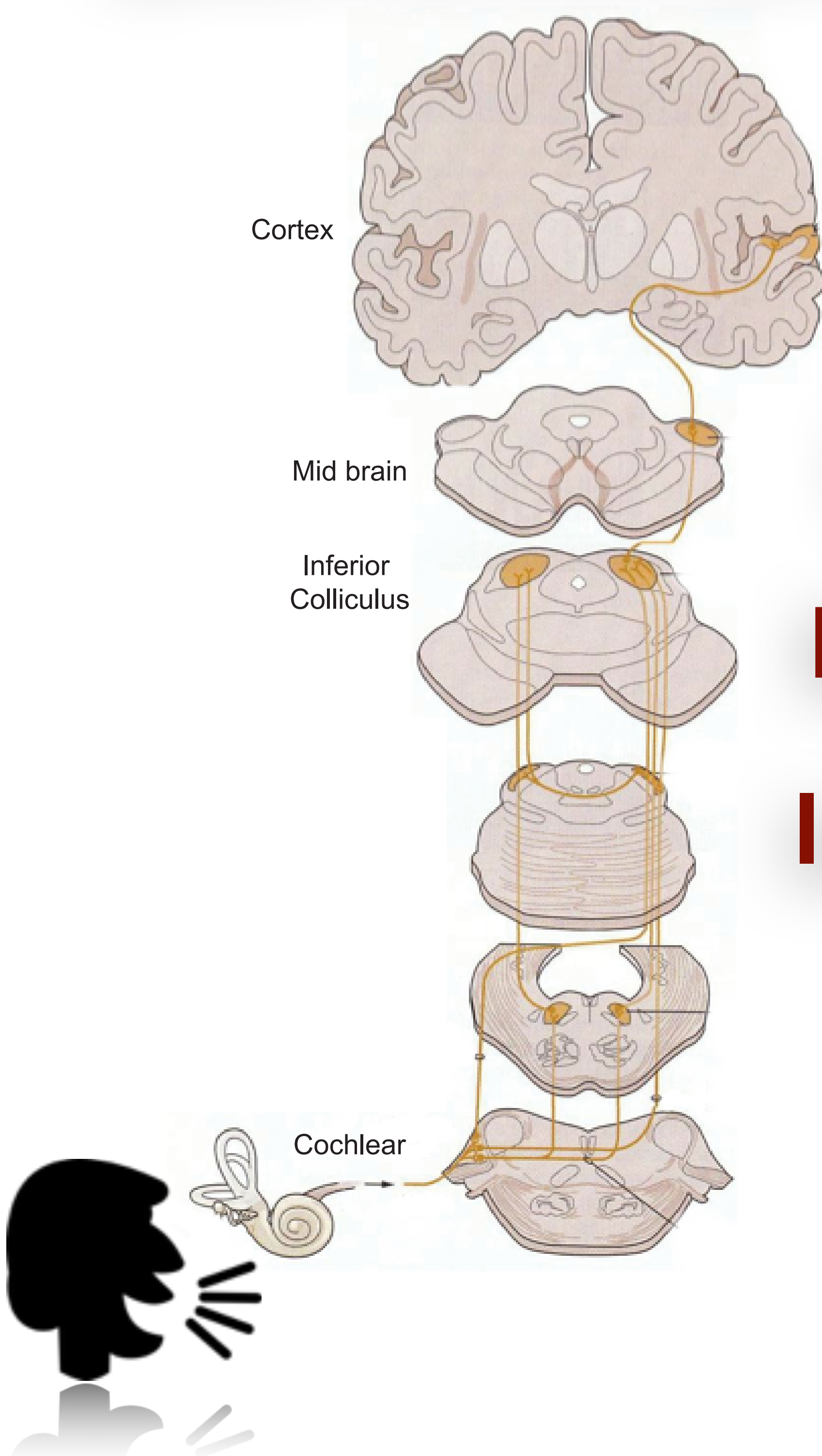
Reduced variability due to **adaptive** mechanisms (e.g. synaptic depression)

Top-down (e.g. attention) **dynamically** modulate the representation

Creating a model of speech communication

Understanding the brain, speech disorders, prosthesis

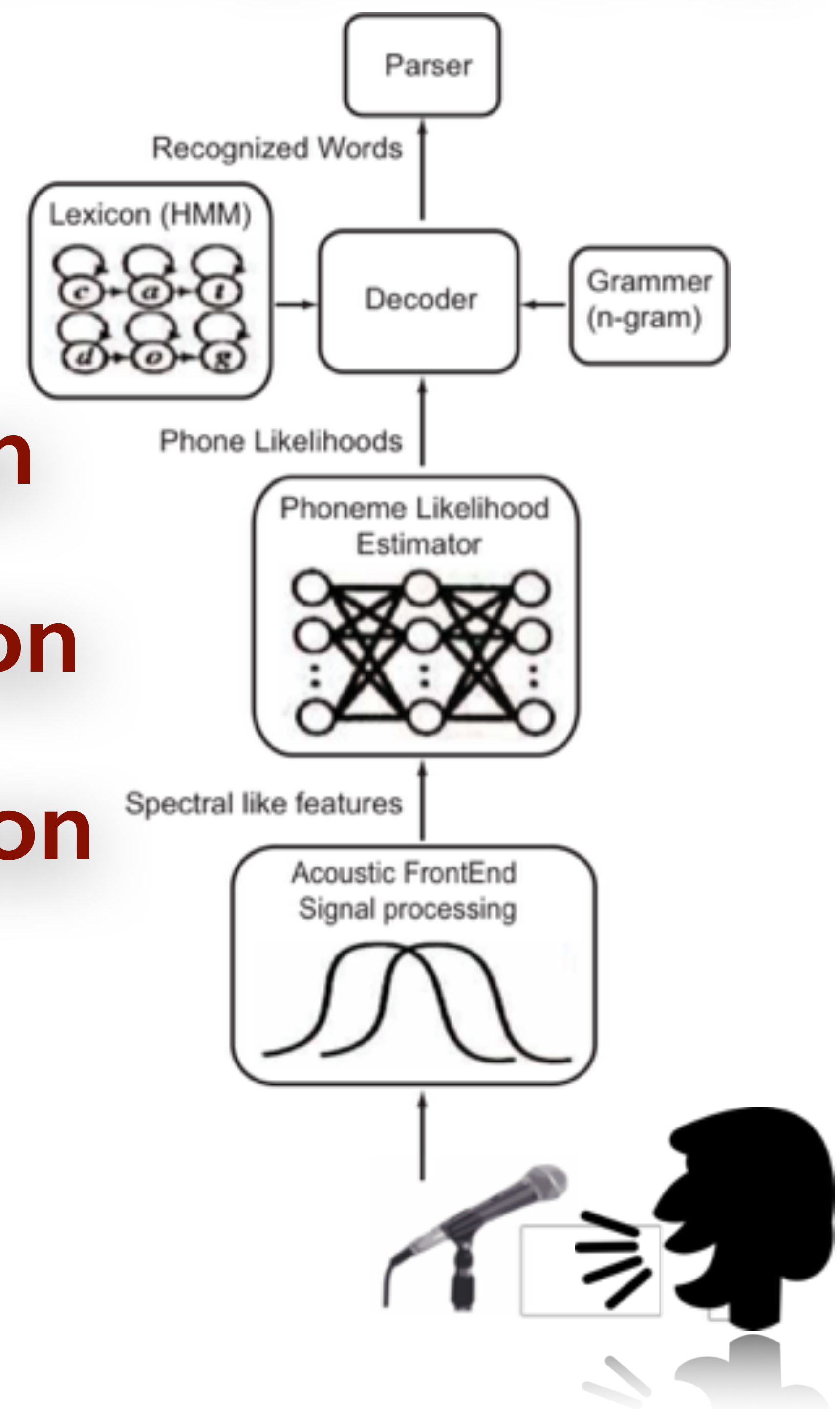
Closing the gap between artificial and biological computing



Computation

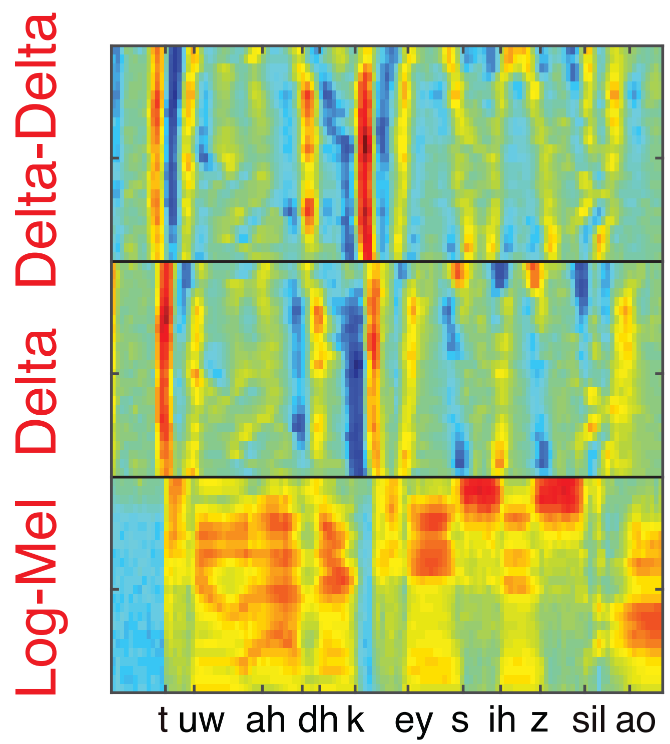
Representation

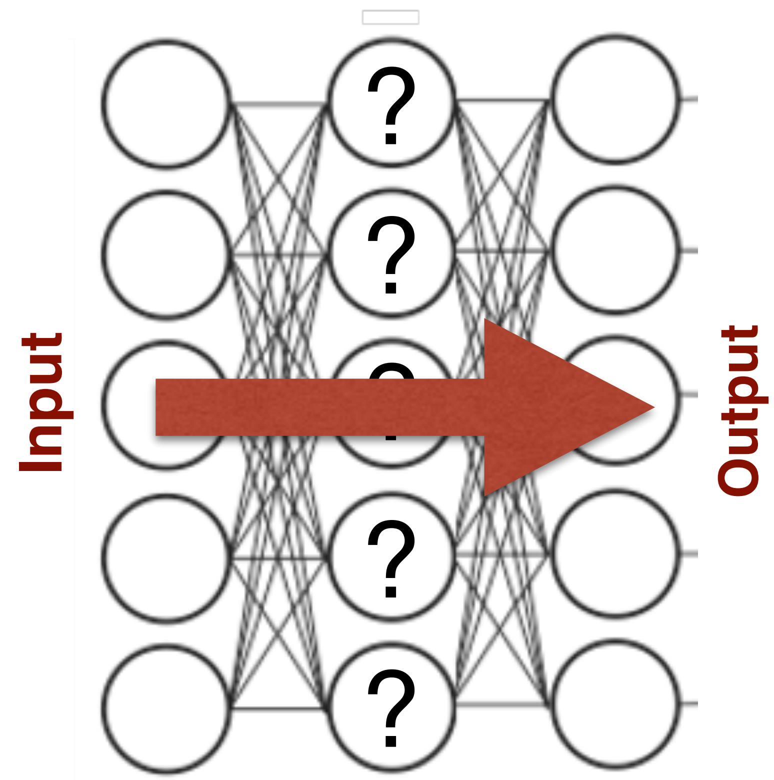
Implementation

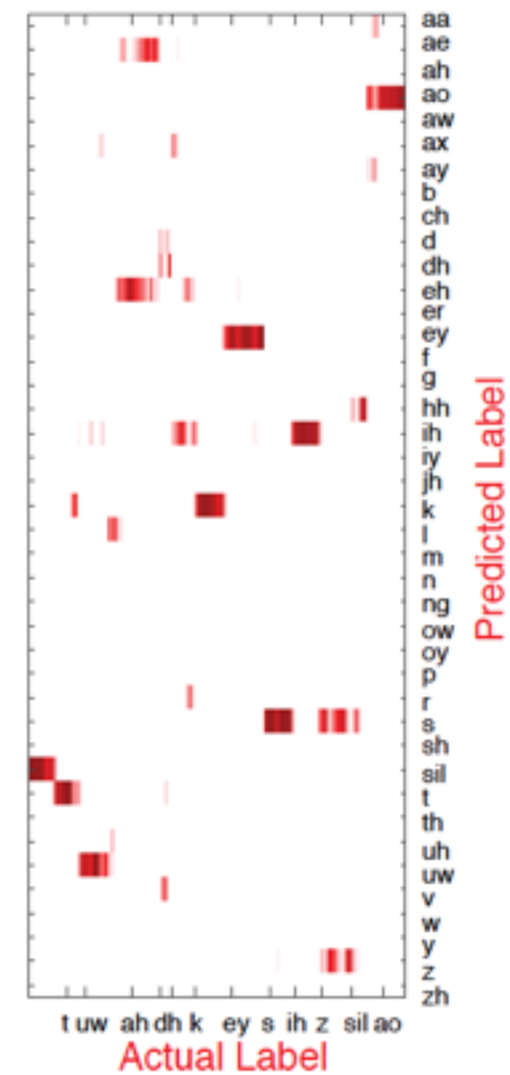


Neuro-inspired models for acoustic modeling

Forming a better understanding of DNN's computation and limitation







What representation is used?

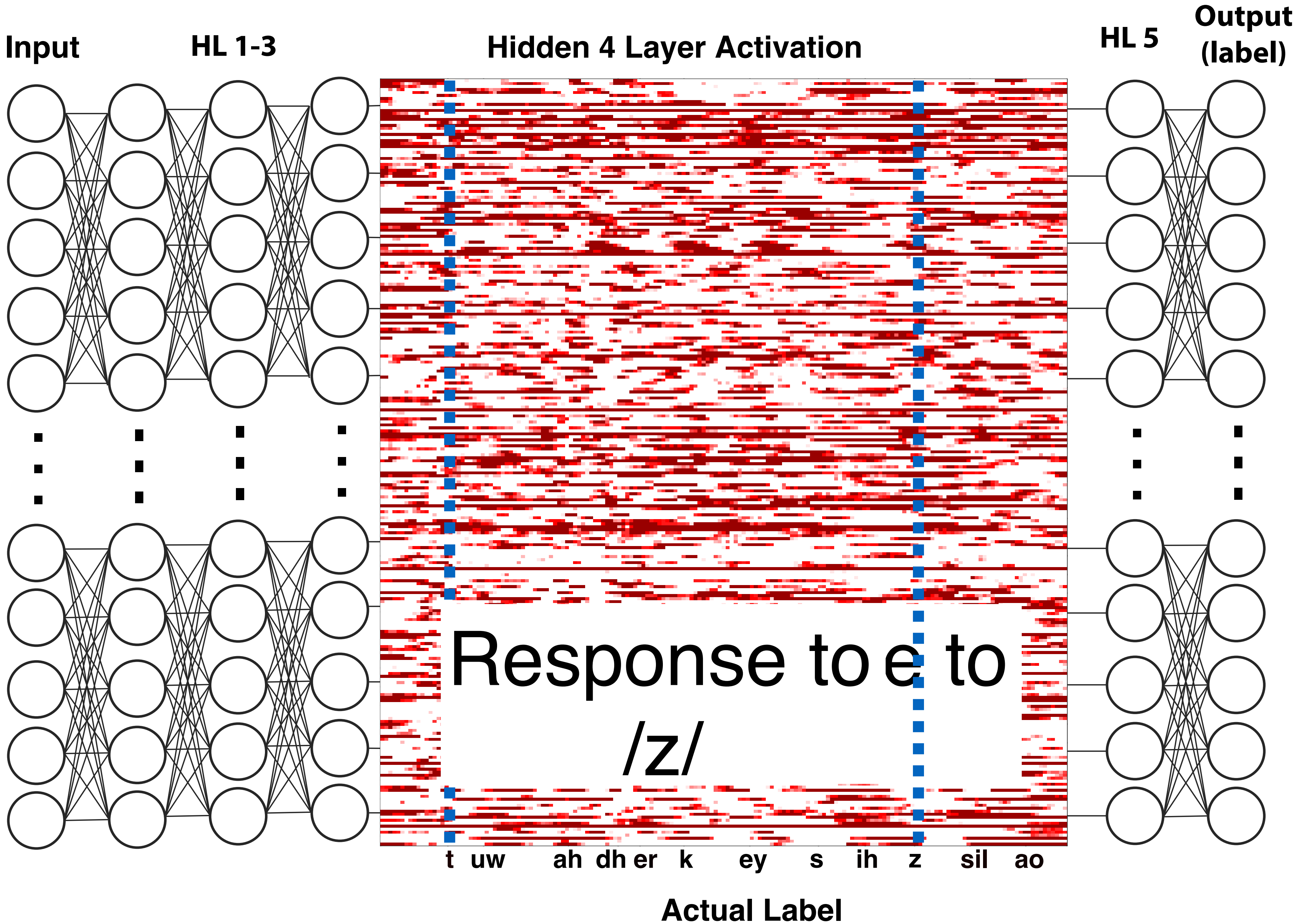
What nonlinear transformation occurs from one layer to next?

Node activations to speech

Input layer, 11 frames of log-mel filterbank, and deltas, trained on WSJ clean

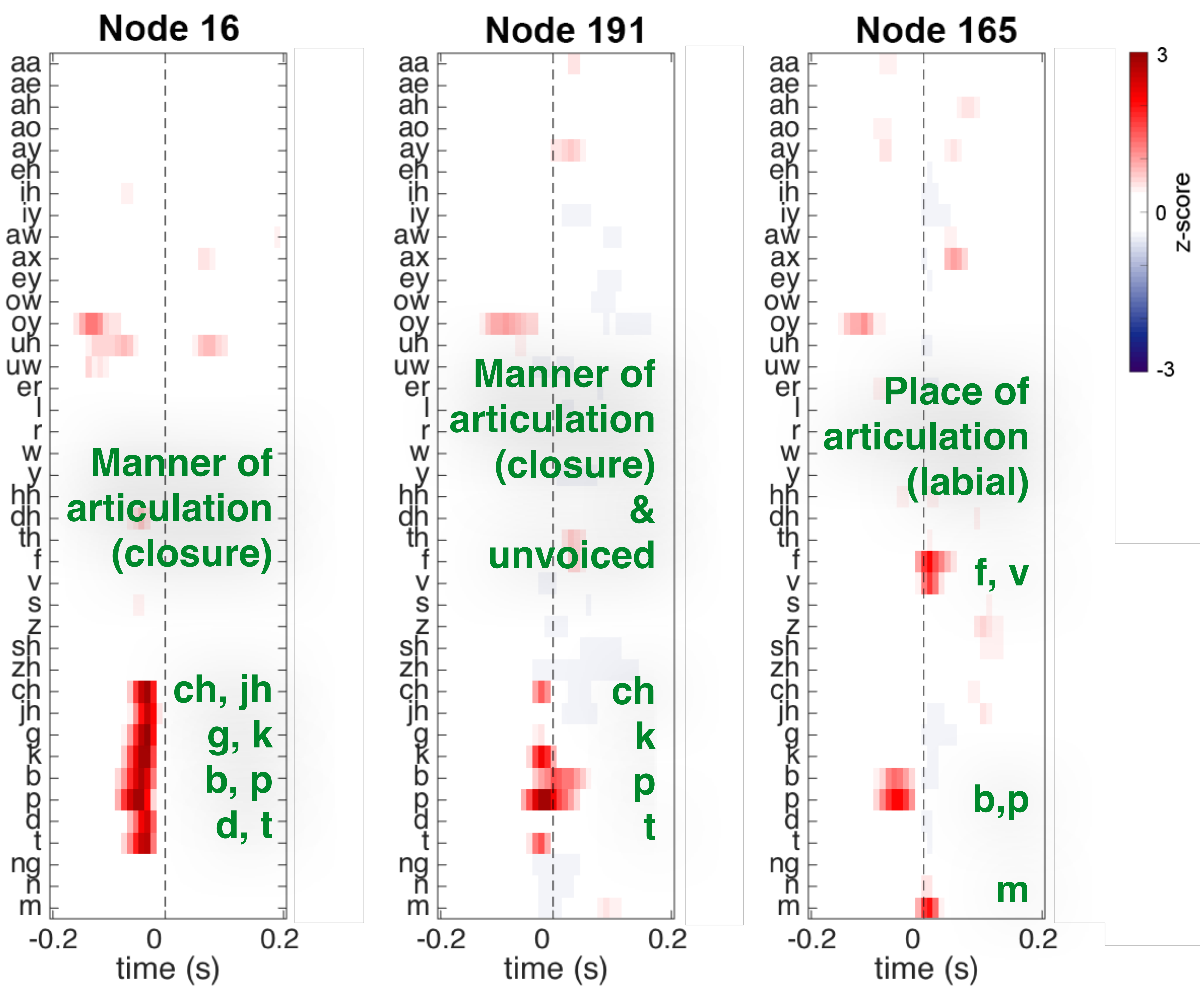
5 sigmoid, hidden layers, 256 nodes each, fully connected, feed-forward

Softmax output 41 nodes context independent

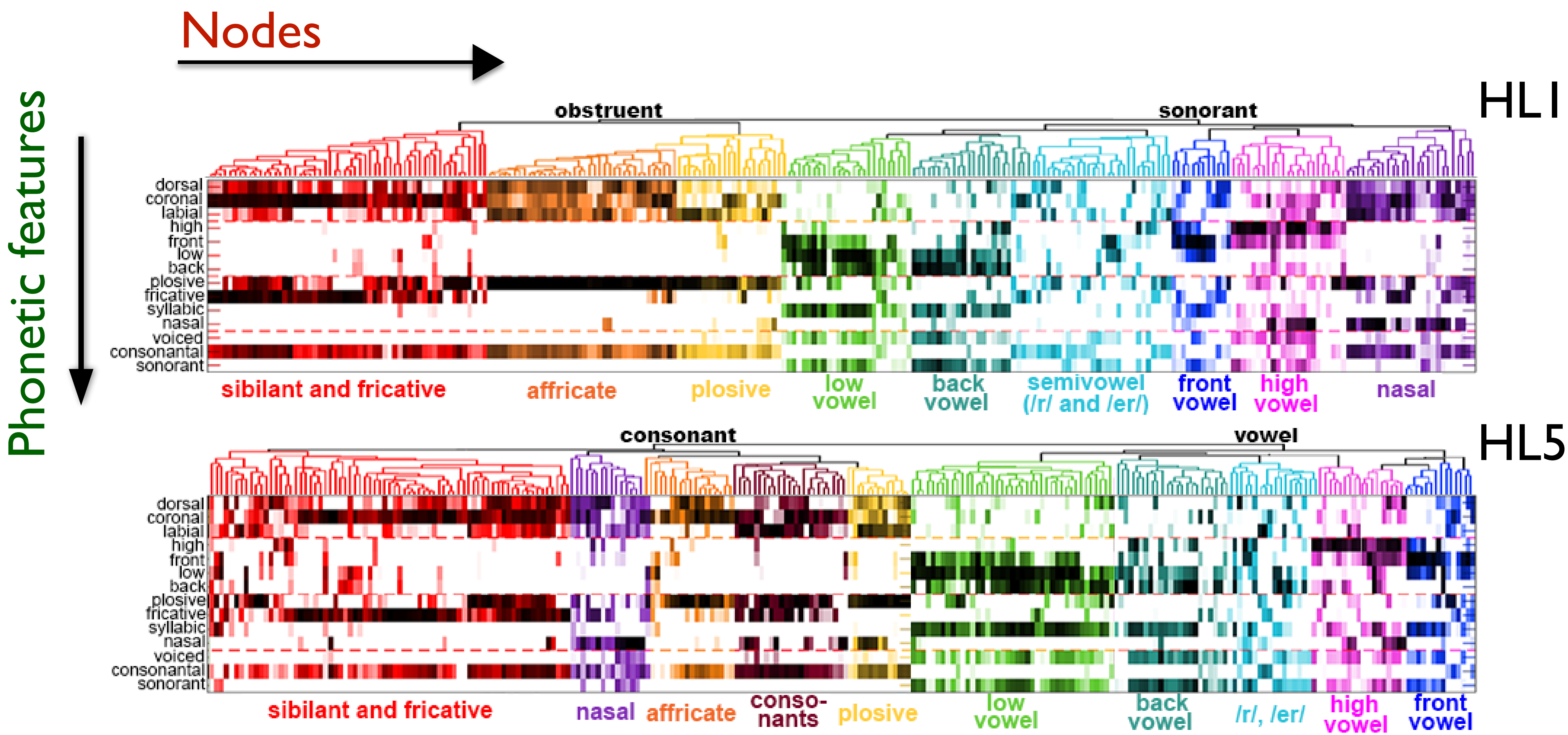


What do the nodes respond to?

Individual nodes become responsive to various phonetic features



What features organize the hidden representations?



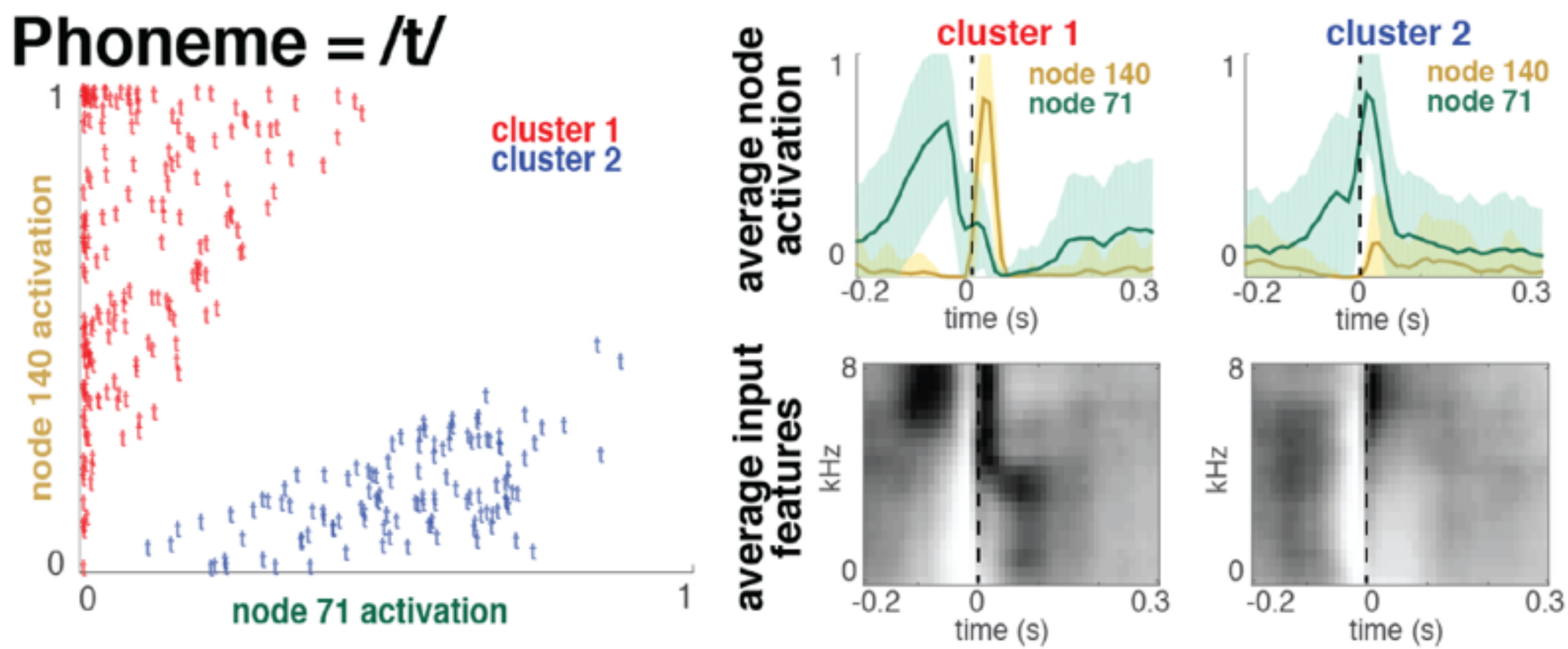
Progressive representation of phonetic features learned by the network that was trained to extract Phonemes from speech

Nagamine et. al. Interspeech 2015

Solving the "invariance problem"?

A phoneme instance (phone) is affected by speaker, context, mood, etc., but perception is robust

Clustering "phones" based on the response of nodes selective to same phoneme



Network learns the variability of the phonemes (phones) and models them explicitly with different nodes

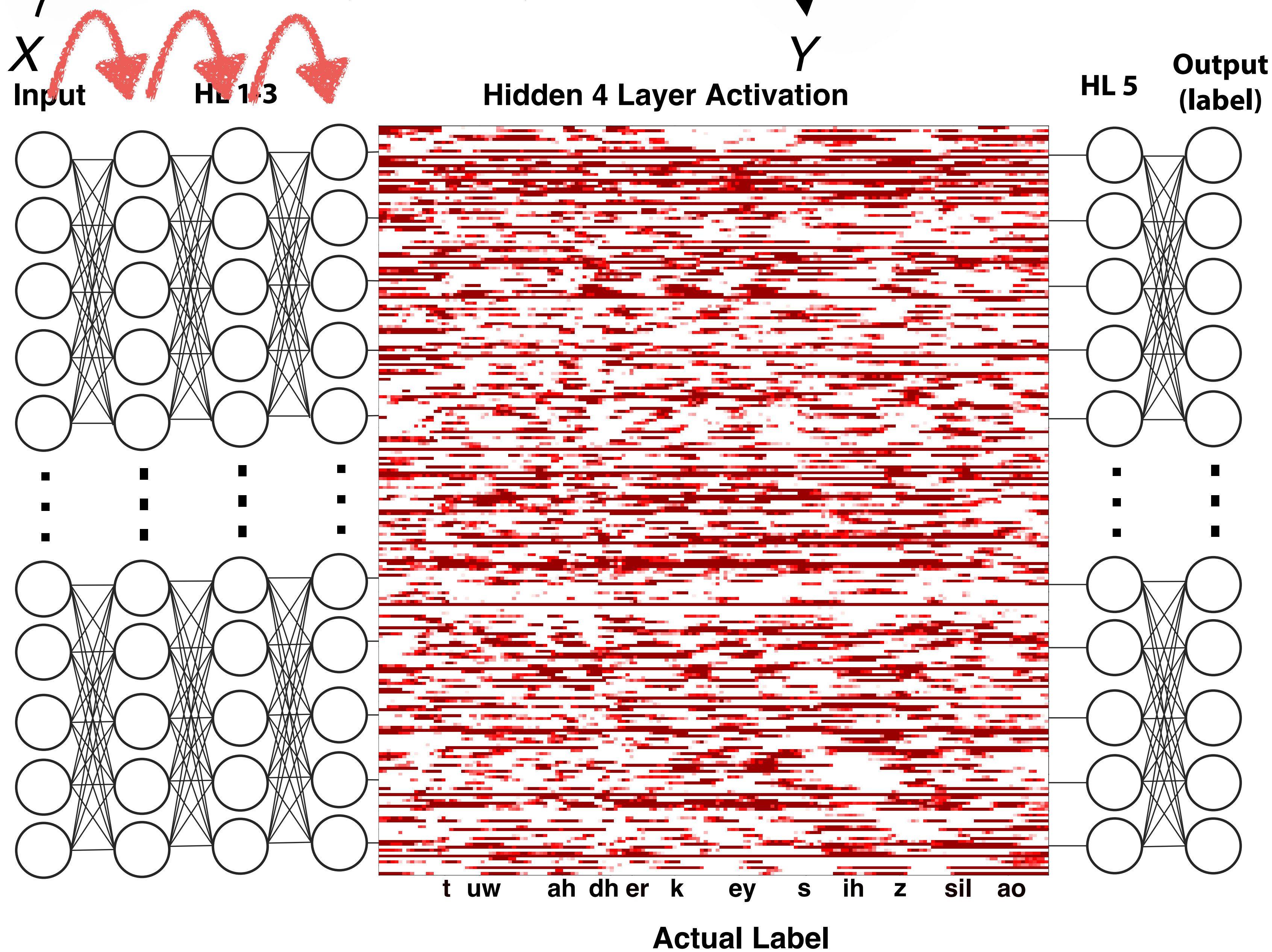
What transformations occurs from layer to layer?

$$\sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

?

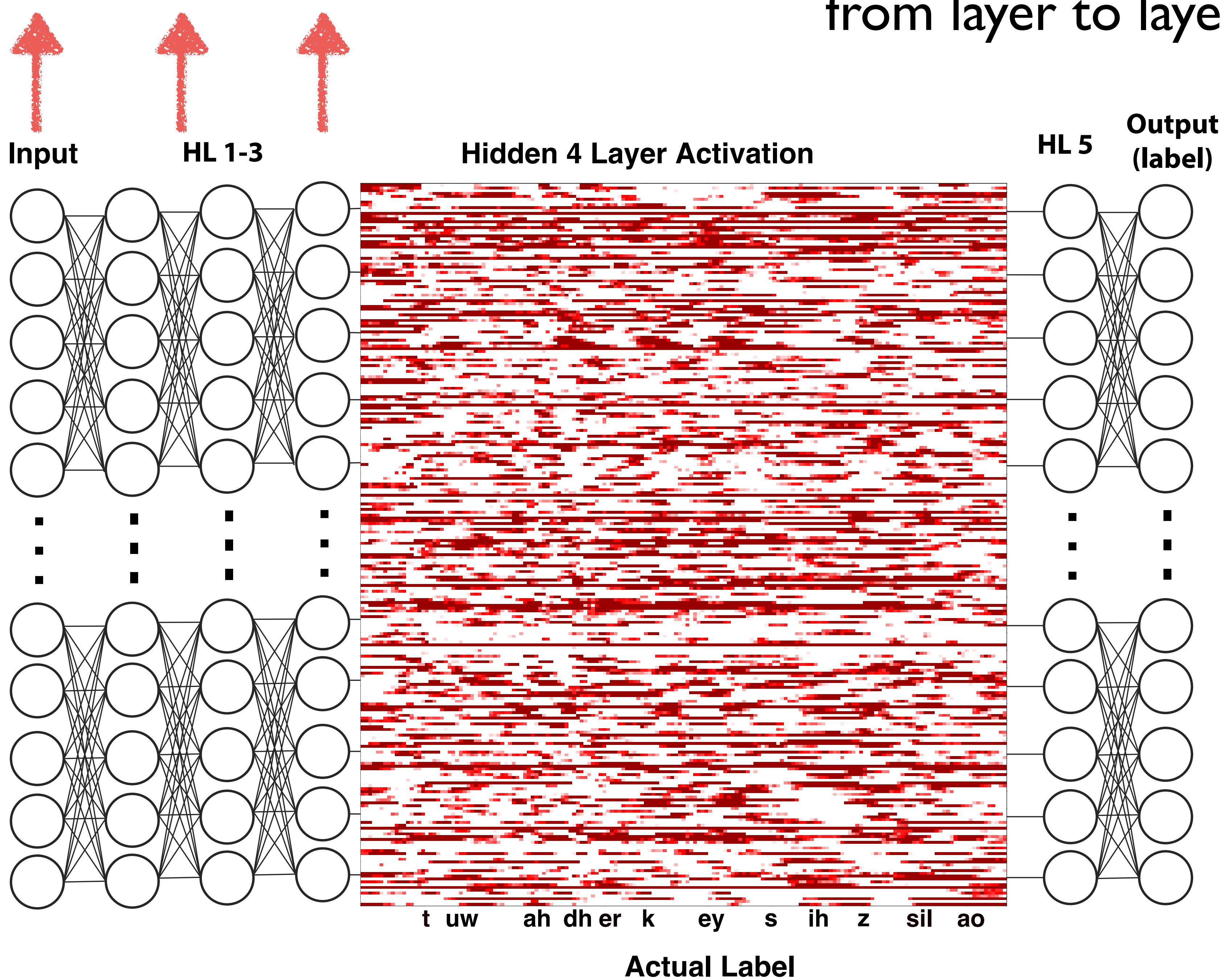
Regression errors are the nonlinear part of the responses



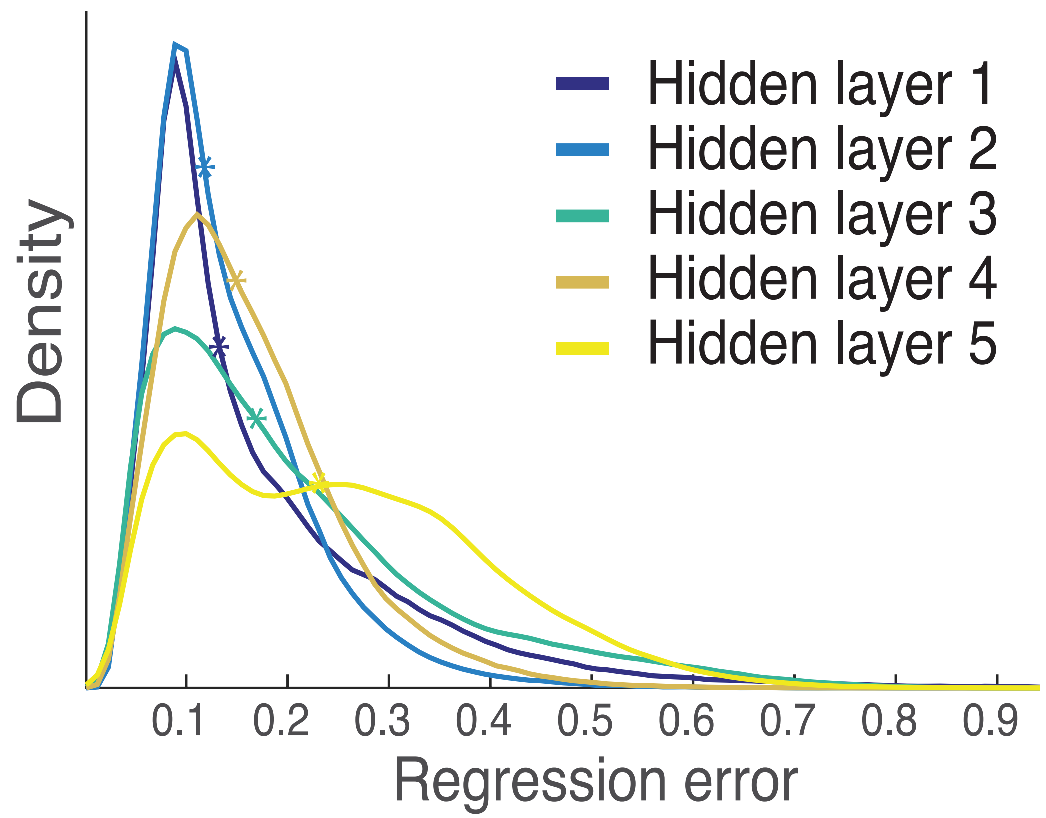
How separable are phonemes in each layer?

Softmax decoding of prob (*Phoneme* | *x*)
Softmax decoding of prob (*Phoneme* | *x*)
Softmax decoding of prob (*Phoneme* | *x*)

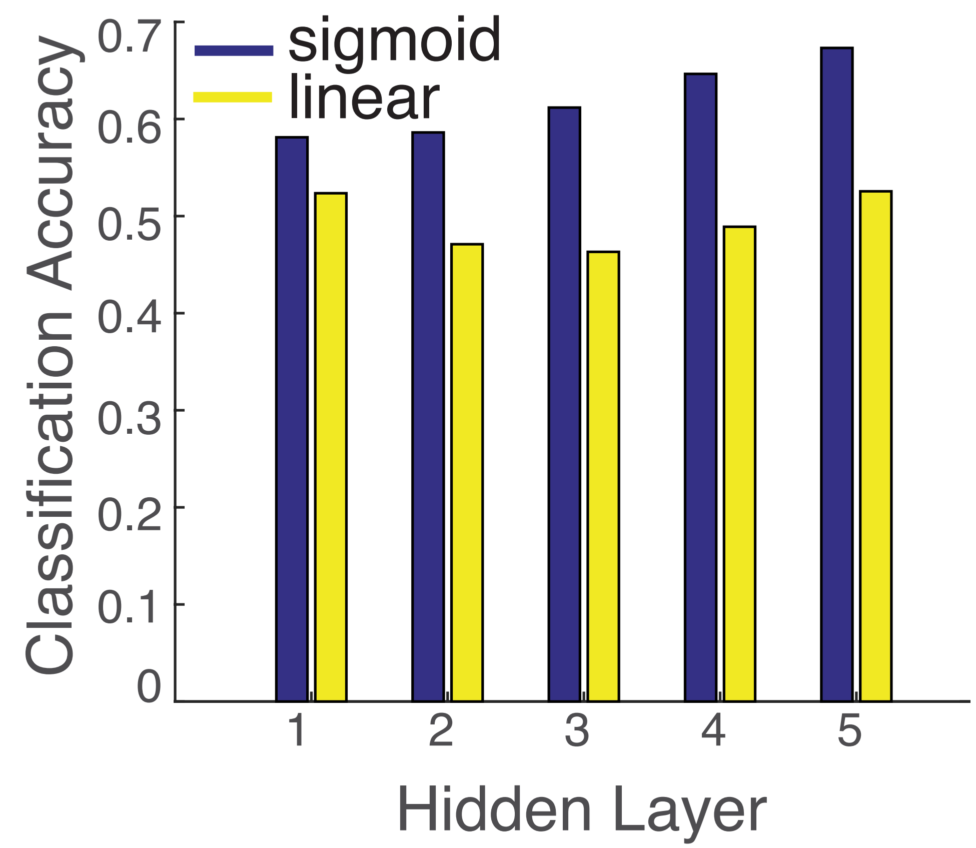
Separability of phoneme categories from layer to layer



The representation becomes increasingly nonlinear and separable



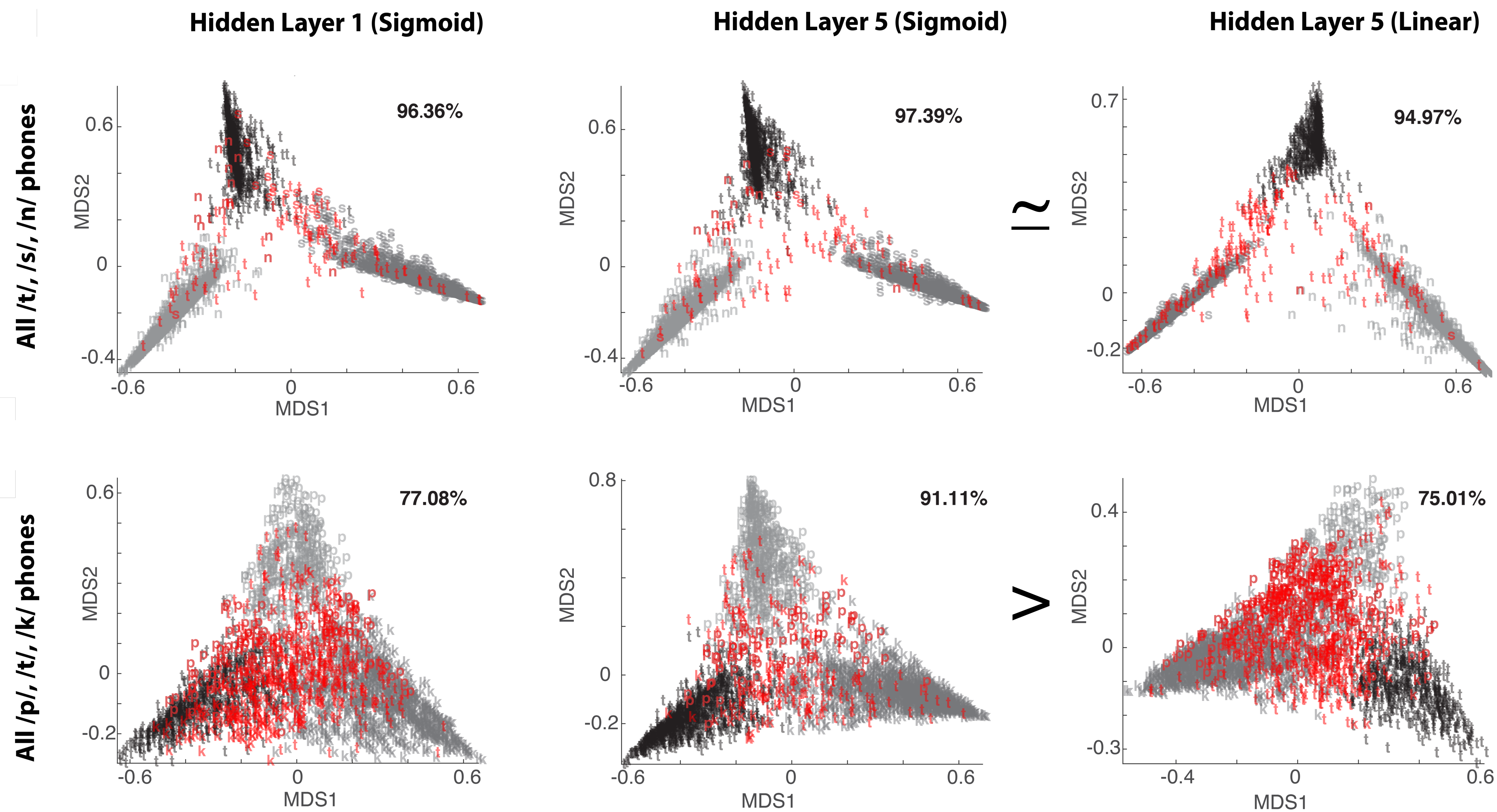
Deeper layers create more separable representation



But *What* becomes more separable, and *How*?

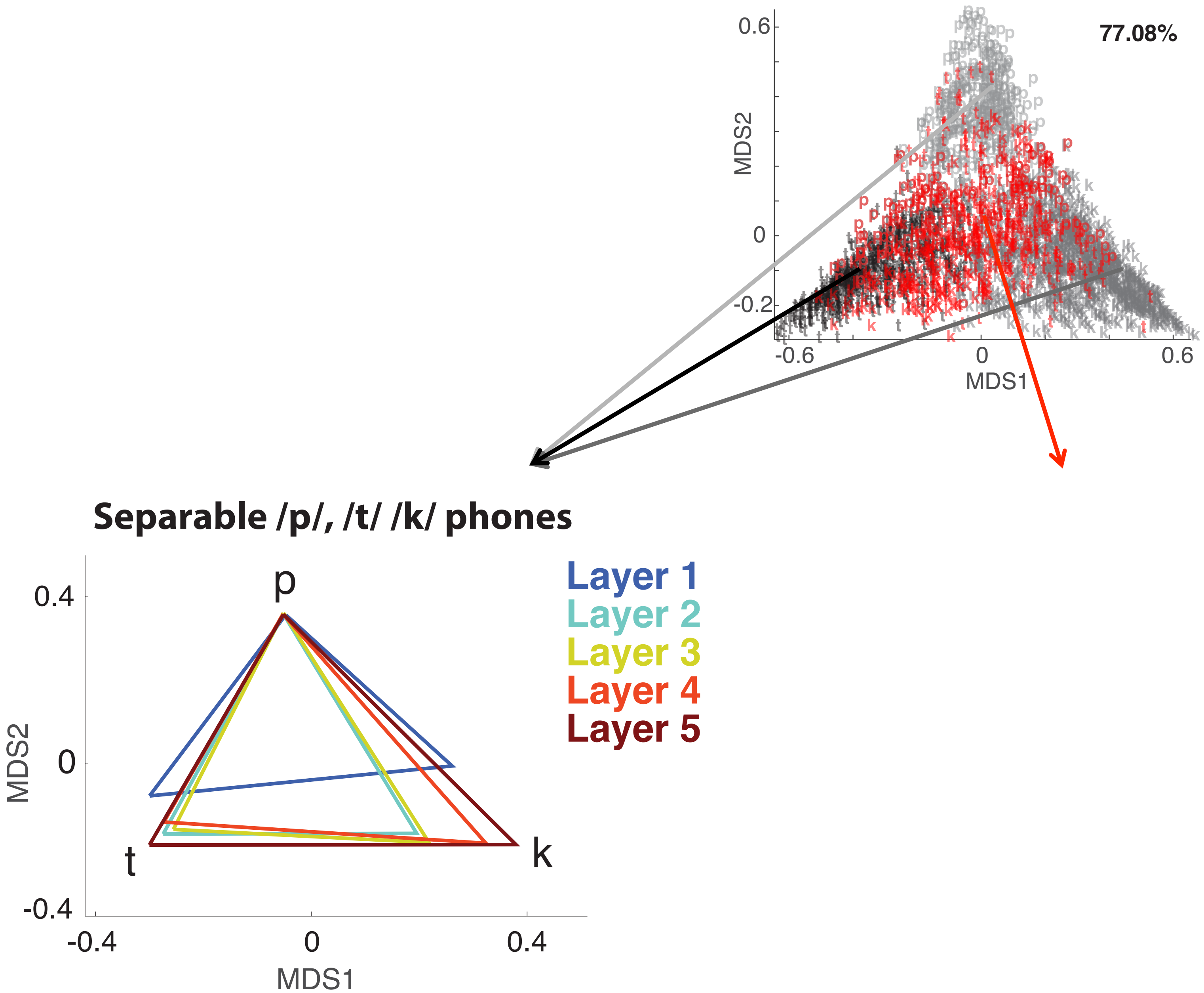
All phoneme/phones or only some?

Decoding phonemes from different layers



Inseparable phones become more separable due to nonlinear transformations

Nonlinear warping of the feature space in the network

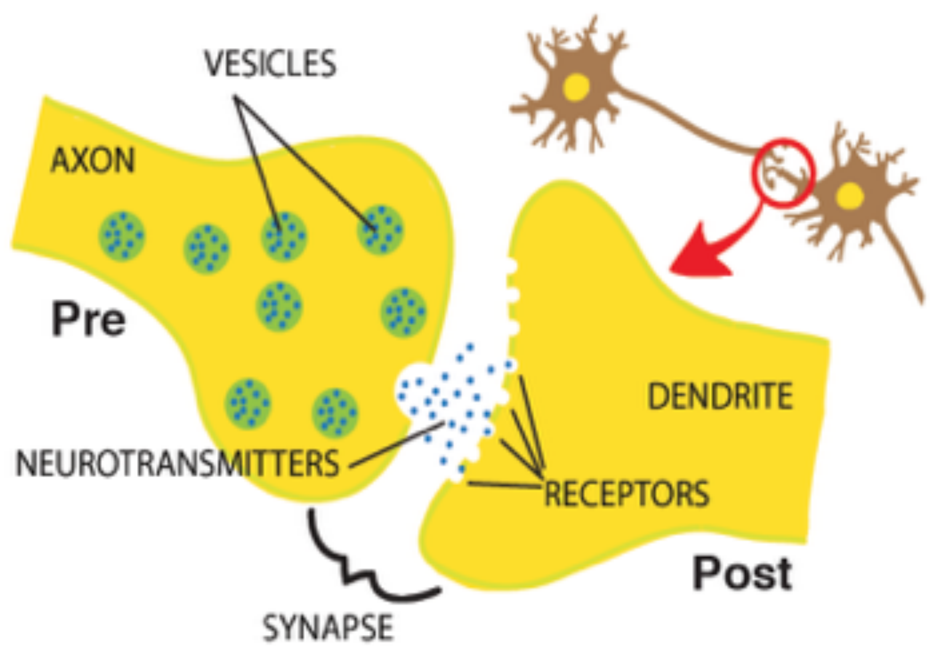


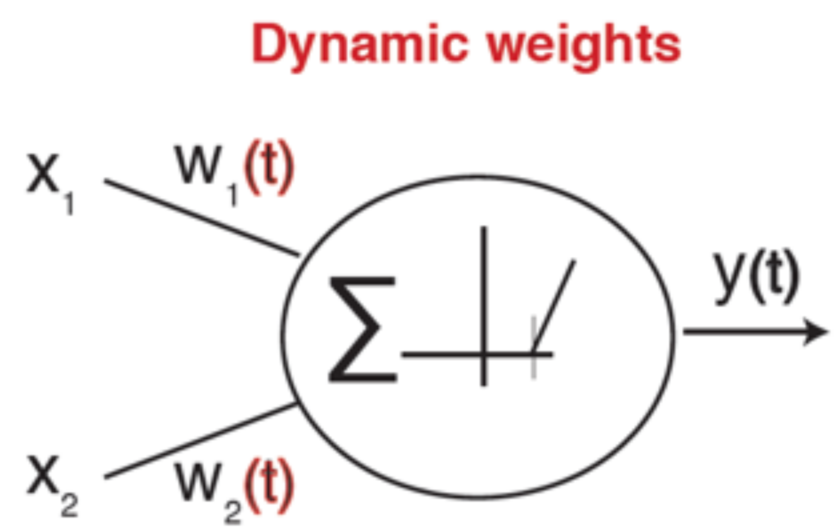
The DNN selectively and progressively **stretches** the feature space to “carve” phonetic categories

Representational properties of DNN

- Progressive selectivity to phonetic features in DNN layers
- Network solves the "invariance problem" by explicitly modeling the sources of variability
- Non-uniform, category-driven nonlinear stretching of acoustic space
- Incorporating neuro-inspired mechanisms?

Synaptic depression in biological neural networks

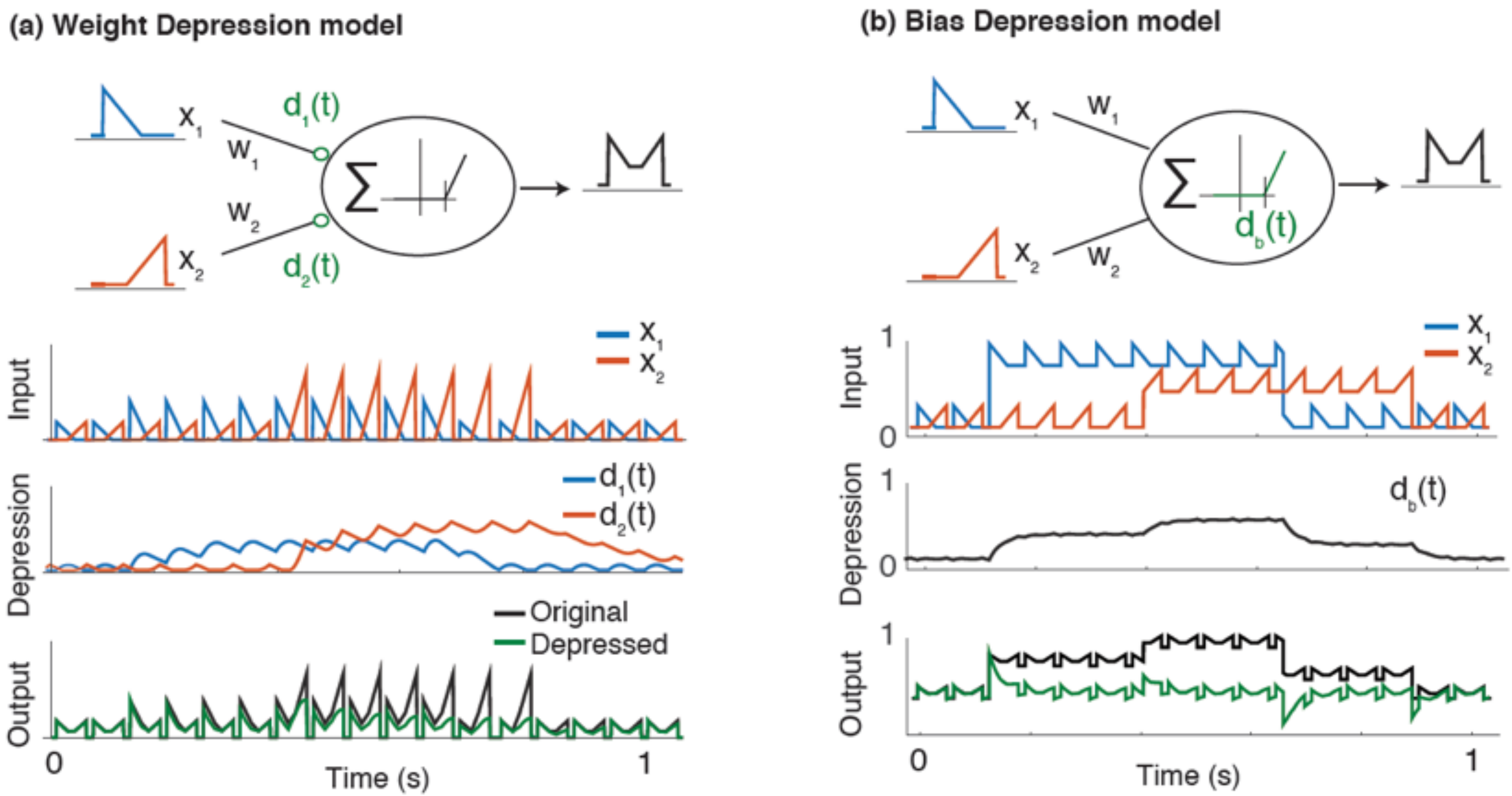




$$z(t) = \sum_i w_i x_i(t) + b$$

$$y(t) = f(z(t))$$

Modeling synaptic depression



$$y(t) = f(\Sigma_i w_i(t) x_i(t)(1 - d_{i,w}(t)) + b)$$

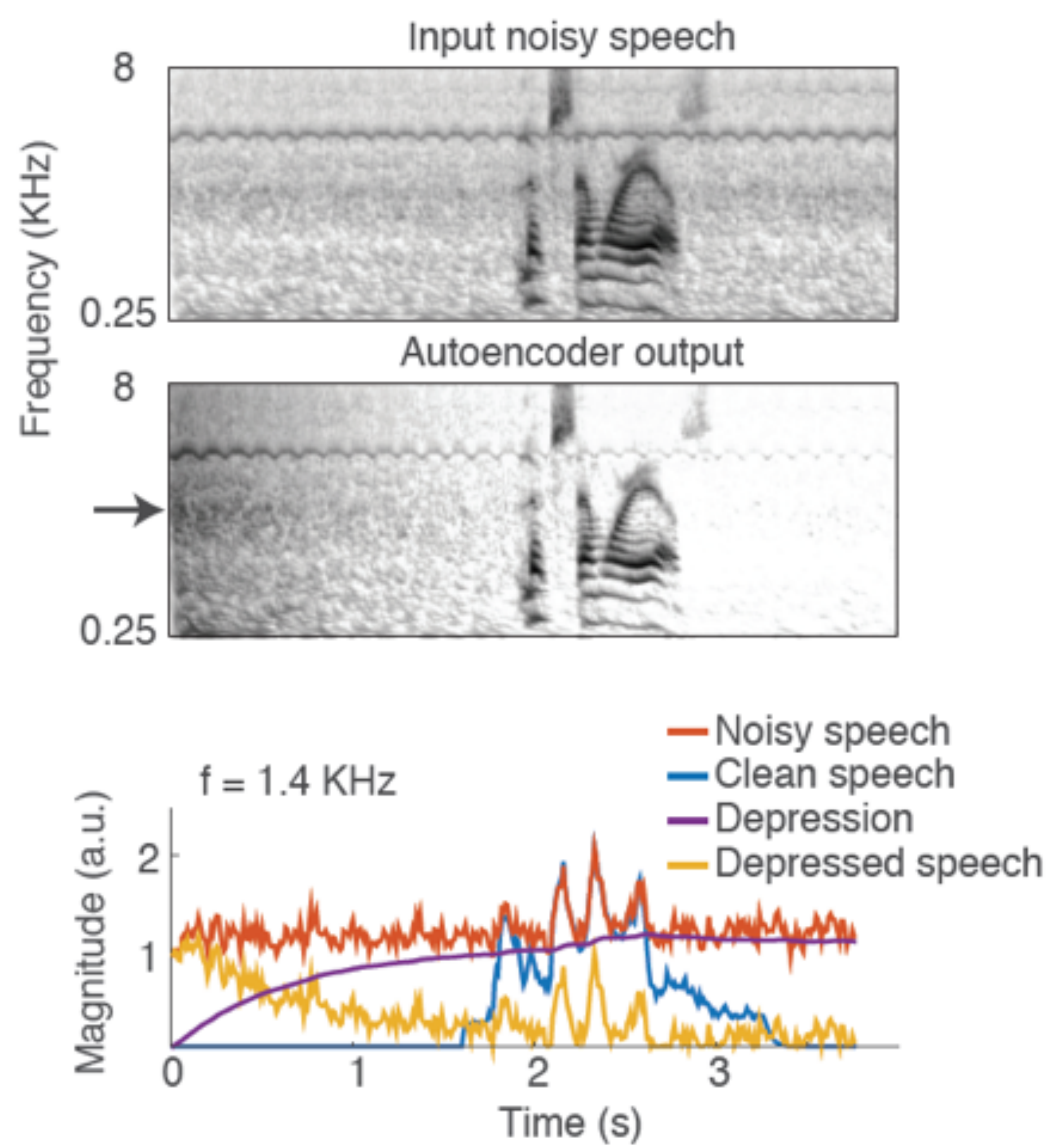
$$d_{i,w}(t) = (1 - \frac{1}{\tau}) d_{i,w}(t-1) + v x_i(t-1)(1 - d_{i,w}(t-1))$$

$$y(t) = f(\Sigma_i w_i x_i + b - d_b(t))$$

$$d_b(t) = (1 - \frac{1}{\tau}) d_b(t-1) + v z_0(t-1)$$

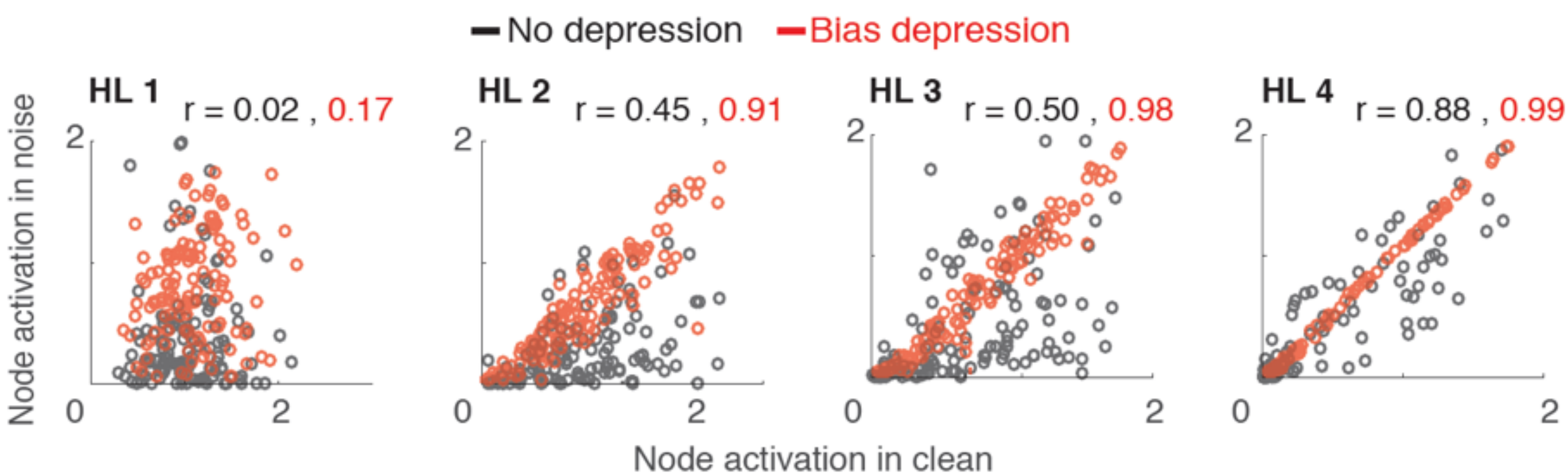
Adaptive, nonlinear effects of synaptic depression

Autoencoder network with/without SD



Bias depression in a DNN for phoneme recognition

Synaptic depression stabilizes the average activation of nodes in noise conditions



Synaptic depression in DNN for phoneme recognition

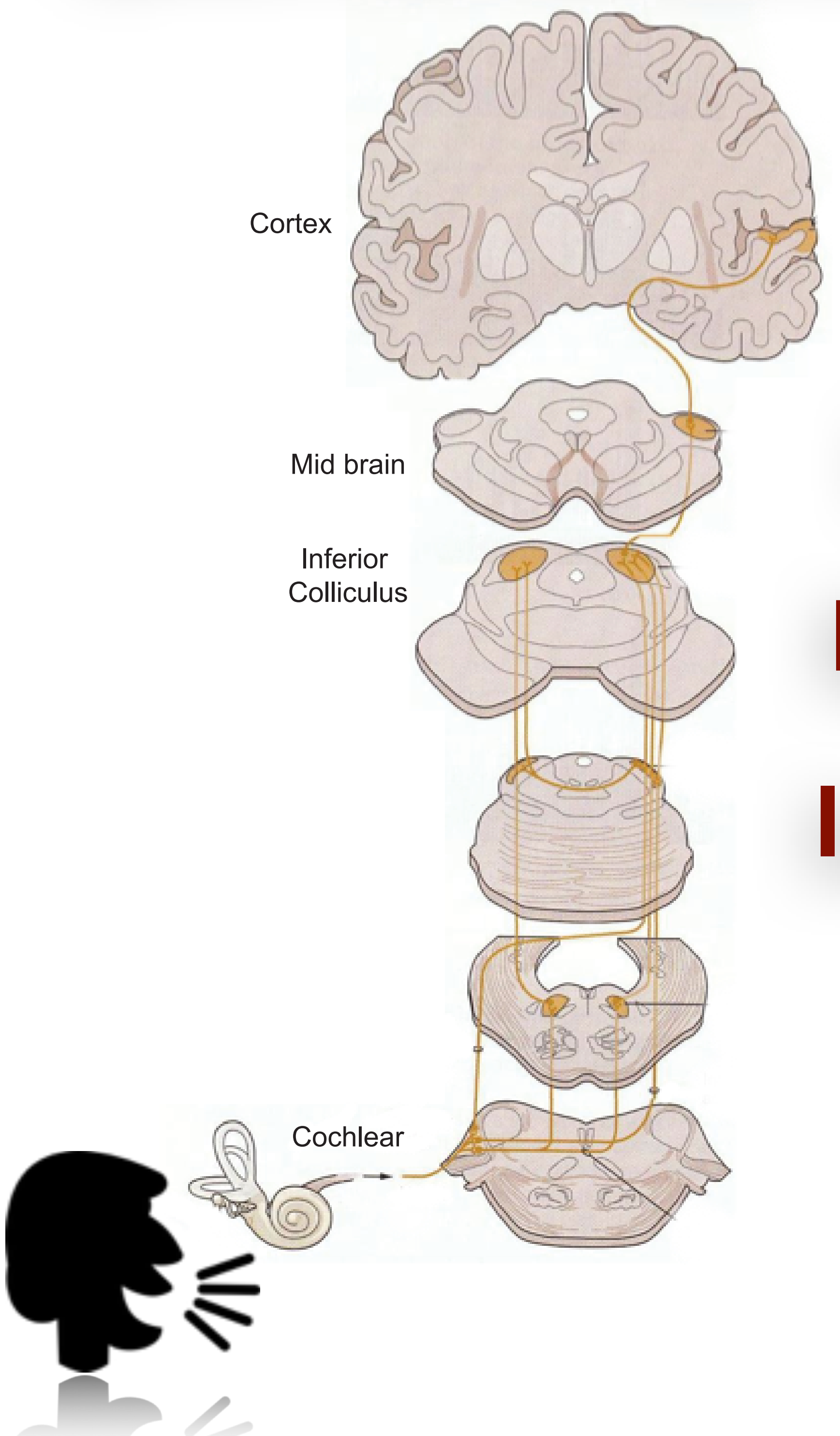
Phoneme classification accuracy: without depression / with depression

SNR	White Noise	Pink Noise	Jet Noise	City Noise	Average
INF	56.65% / 54.61%	–	–	–	–
20	35.81% / 41.39%	42.89% / 45.33%	42.13% / 43.83%	47.07% / 46.84%	41.98% / 44.35%
15	27.39% / 34.63%	34.27% / 38.99%	33.13% / 37.18%	41.03% / 42.22%	33.96% / 38.26%
10	19.05% / 26.71%	24.83% / 31.69%	24.28% / 29.23%	33.11% / 35.52%	25.32% / 30.79%
5	12.38% / 19.37%	15.89% / 22.52%	15.48% / 20.94%	23.40% / 27.92%	16.79% / 22.69%
0	7.90% / 14.17%	9.15% / 14.83%	8.45% / 13.85%	14.95% / 19.52%	10.11% / 15.59%
Average	20.51% / 27.25%	25.41% / 30.67%	24.69% / 29.01%	31.91% / 34.40%	–

Creating a model of speech communication

Understanding the brain, speech disorders, prosthesis

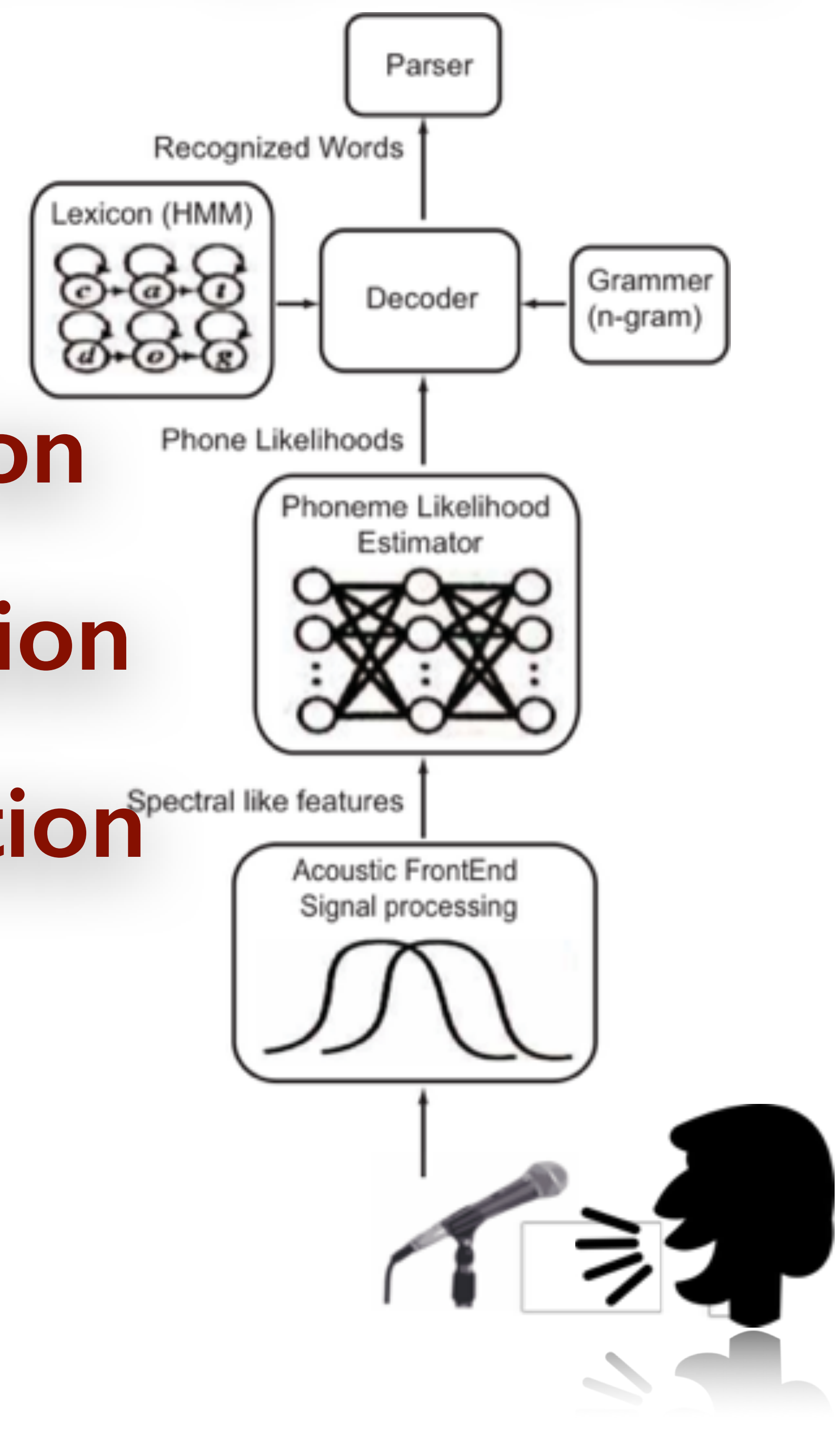
Closing the gap between artificial and biological computing



Computation

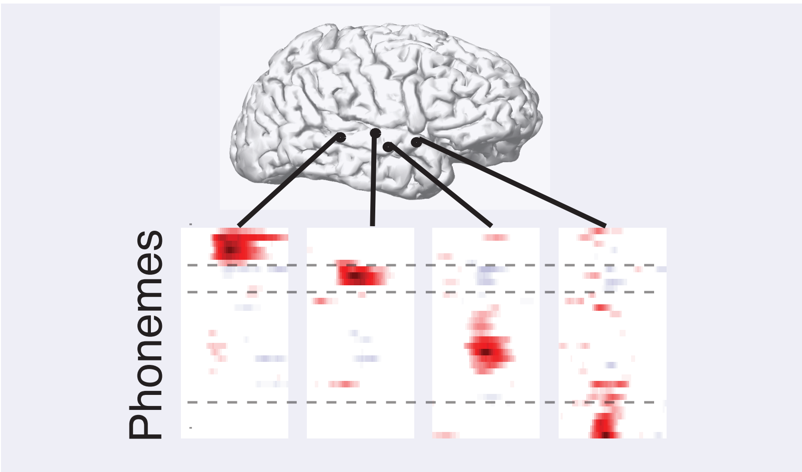
Representation

Implementation

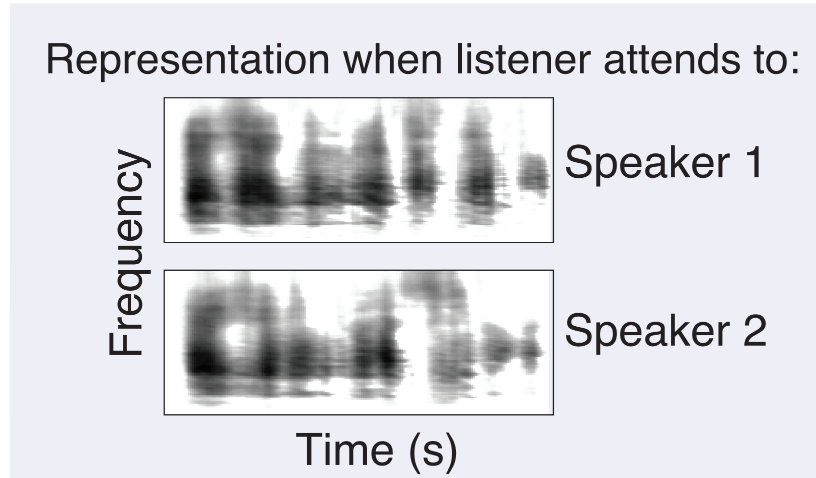


Properties of the cortical representation

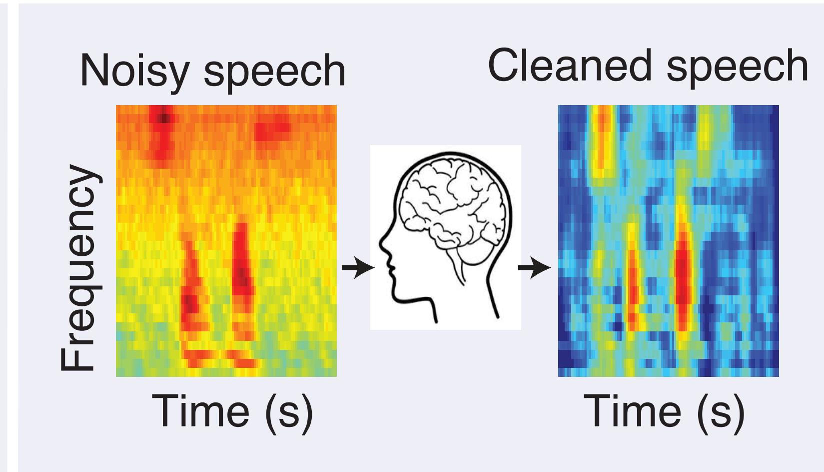
Selective



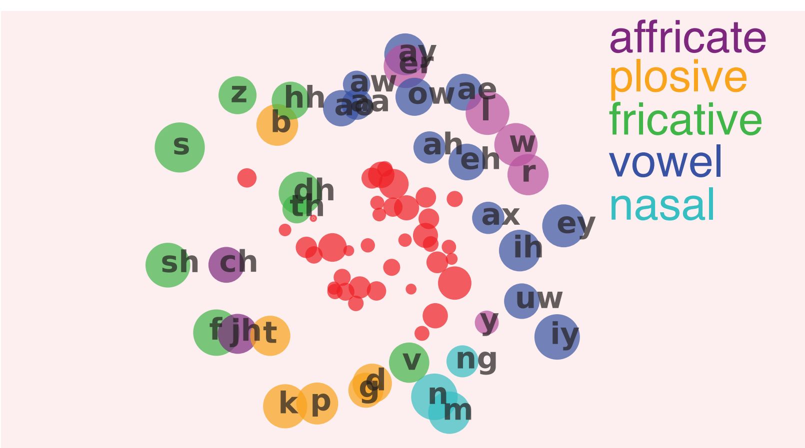
Dynamic



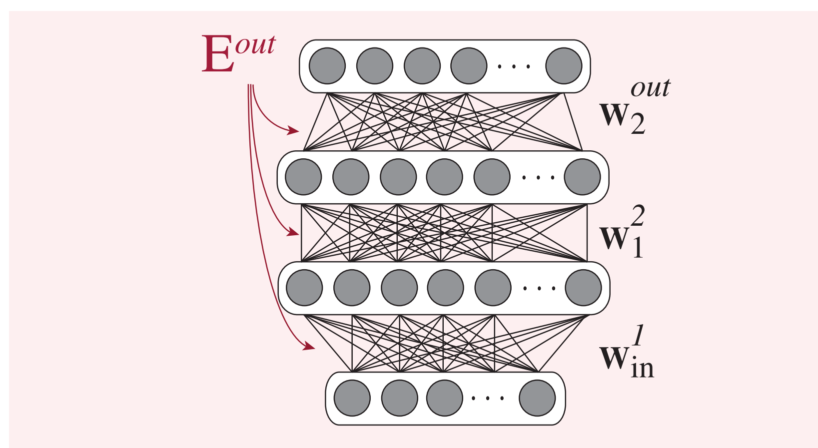
Adaptive



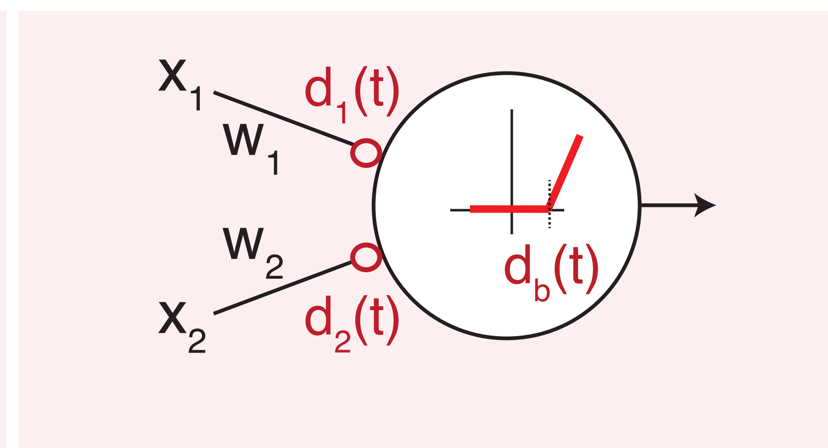
Making new models:



(Nagamine et. al. 2016)



(Nagamine et. al. 2016)



(Zhang et. al., 2016)

Temporal integration, higher order units

What does the feedback change?

Interaction of top-down and bottom-up

Our Lab:

James O'Sullivan

Tasha Nagamine

Laura Long

Zhou Chen

Bahar Khalighinejad

NSF CAREER Award

NIH National Institute on Deafness and Other Communication Disorders

R01DC014279

THE PEW CHARITABLE TRUSTS