On the Formation of Phoneme Categories in DNN Acoustic Models

Tasha Nagamine
Department of Electrical Engineering, Columbia University

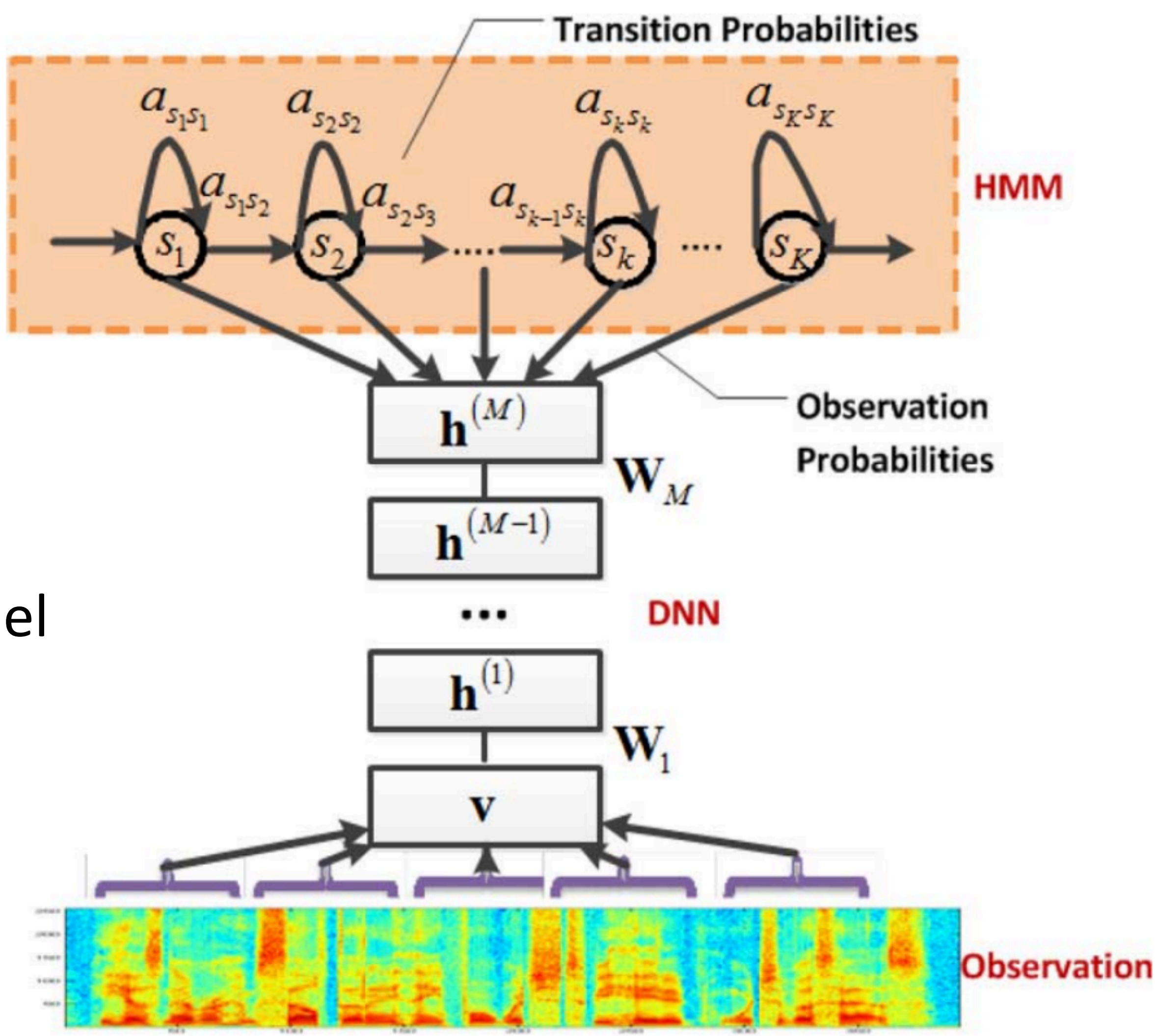
October 14, 2015

Motivation

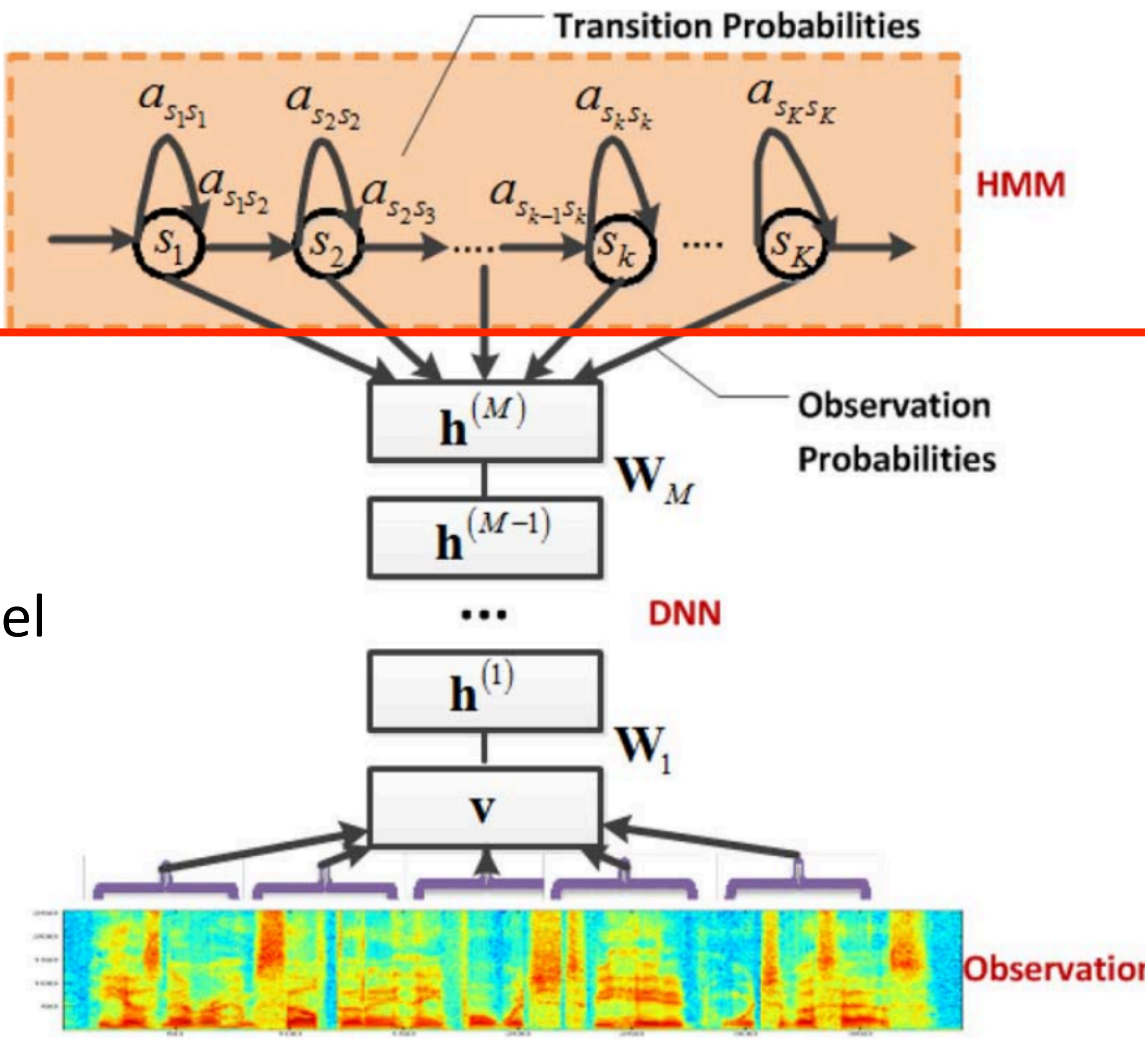
- Large performance gap between humans and state-of-the-art ASR systems
- Computational principles of DNNs remain elusive; they are analytically intractable
- Improving these models requires a better understanding of their transformations

Introduction to acoustic models

acoustic model



Introduction to acoustic models



acoustic model

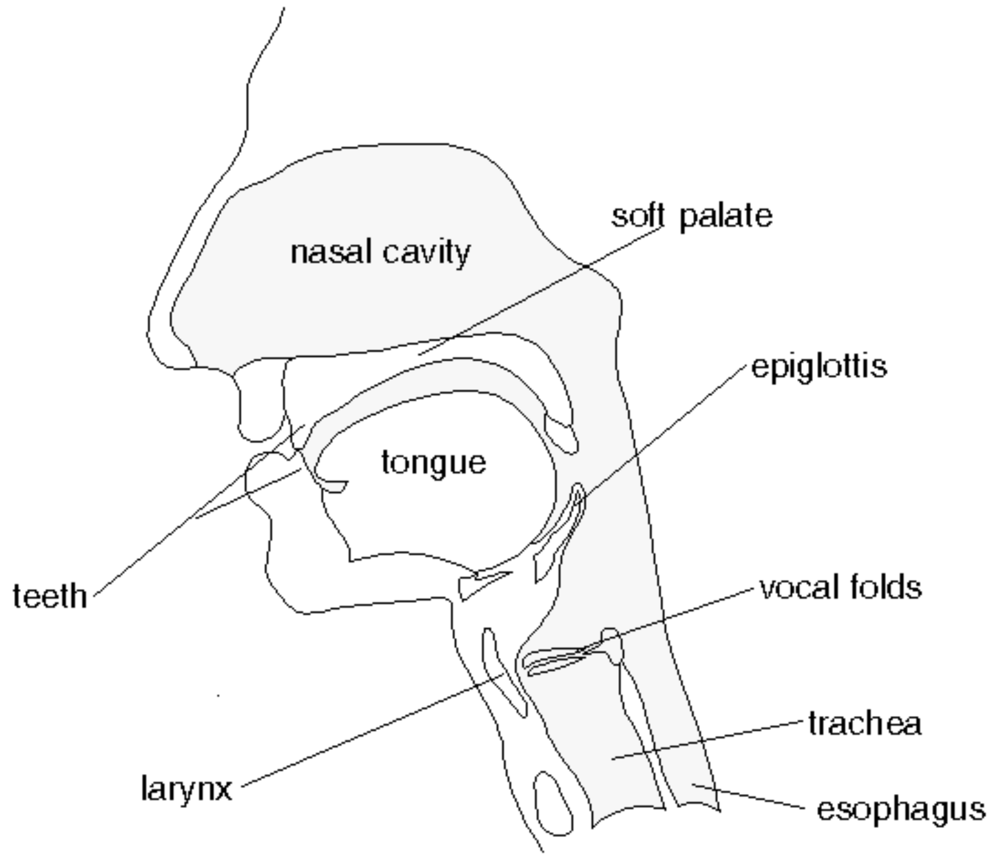
Dahl et al., IEEE Transactions on Audio, Speech, and Signal Processing 2012

Phonemes

Smallest contrastive unit in language

- e.g., “k” vs. “b” in cat/bat
- ~40-60 in English

Output target in acoustic modeling



Phonetic Features

Manner of articulation

Place of articulation

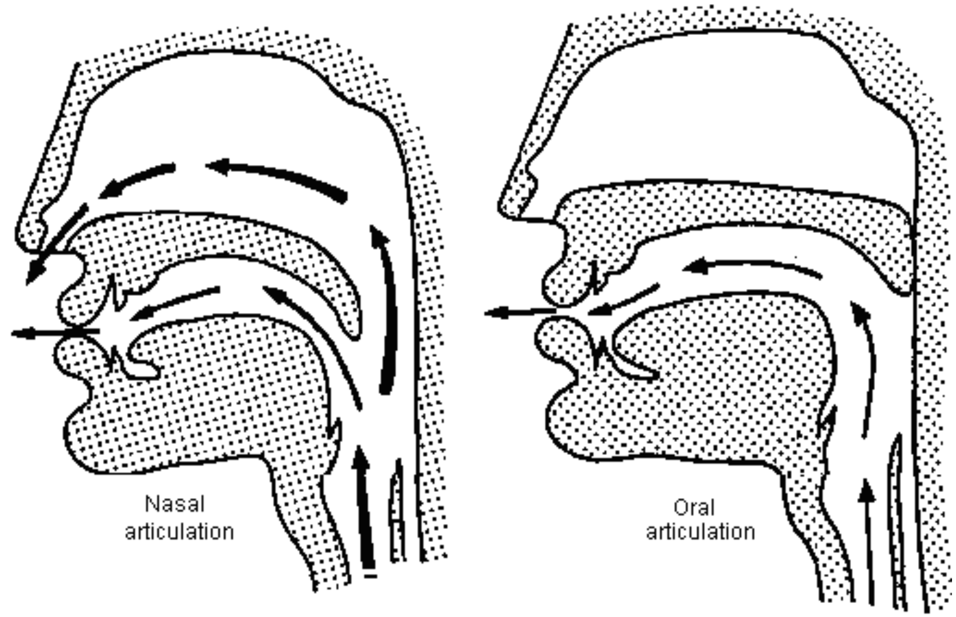
Voicing

Phonetic Features

Manner of articulation

Place of articulation

Voicing

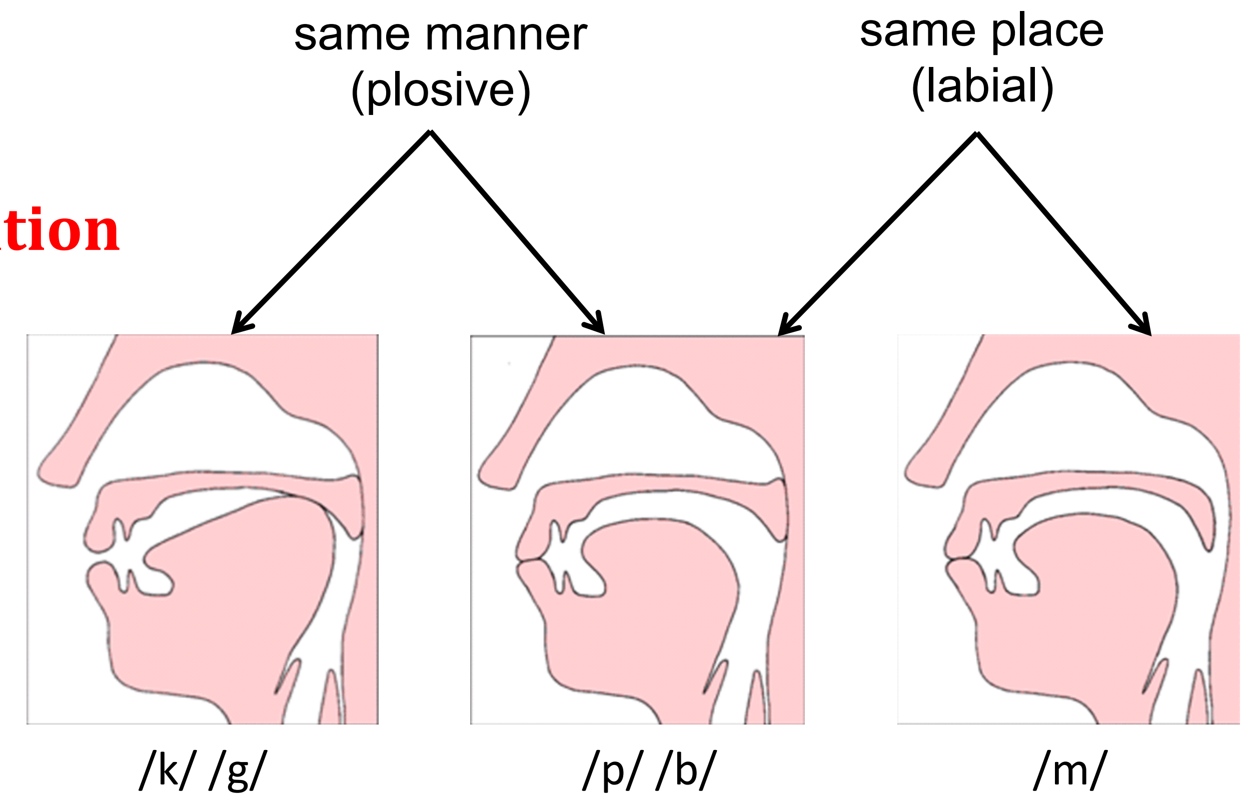


Phonetic Features

Manner of articulation

Place of articulation

Voicing

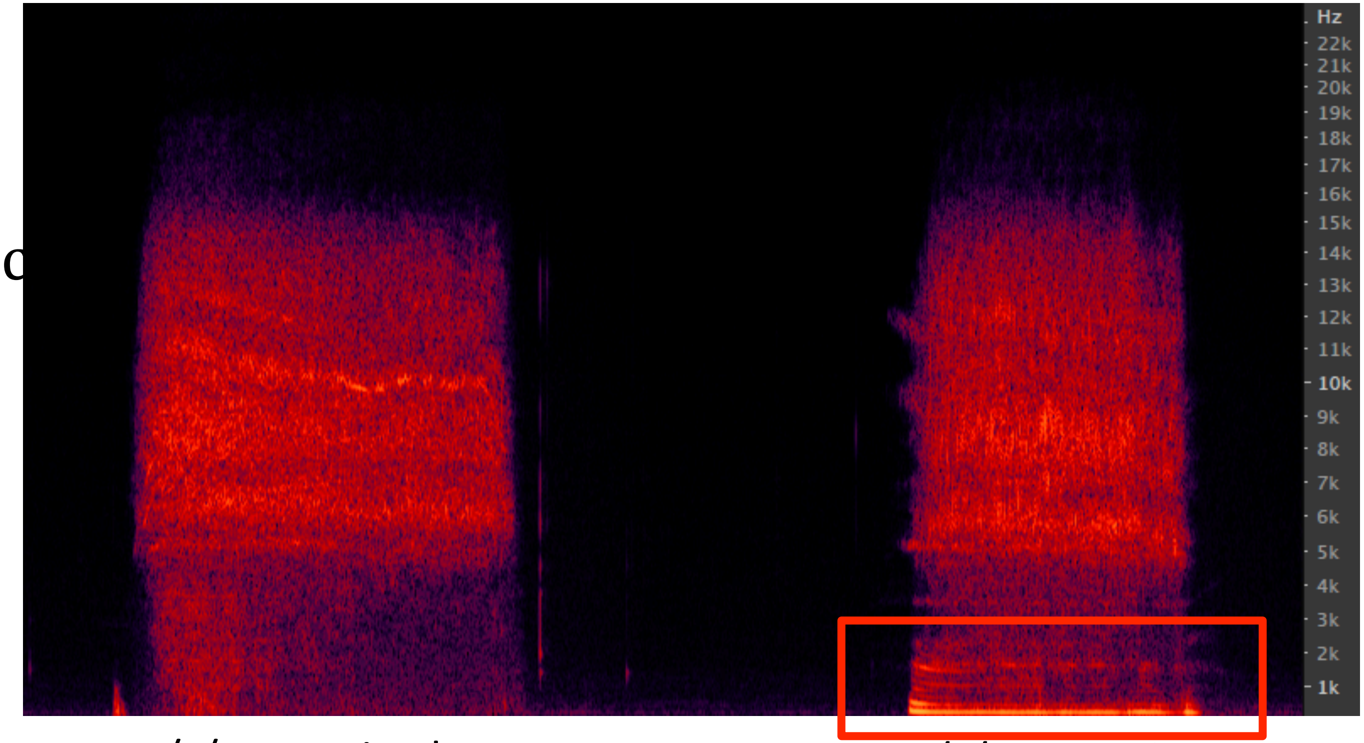


Phonetic Features

Manner of articulation

Place of artic

Voicing

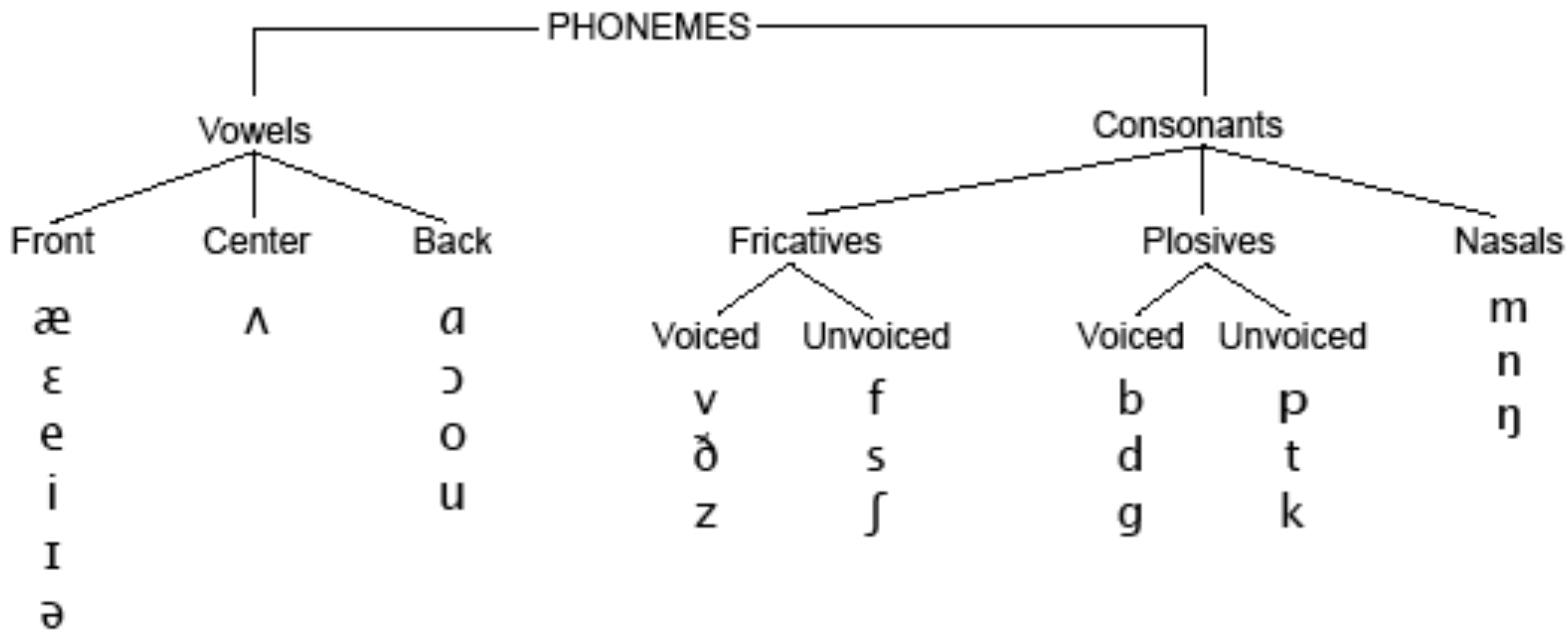


/s/ = unvoiced

/z/ = voiced

same manner (fricative)
same place (alveolar)

Phonetic Features



Distinctive Features
Chomsky, Halle, Stevens

Distinctive Features

Table 1. Distinctive Features of American English Consonants

	p	b	m	f	v	θ	ð	t	d	n	s	z	l	ɾ	ʃ	ʒ	tʃ	dʒ	j	ɹ	k	g	ŋ	w	ʔ	h
Back	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
High	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-
Coronal	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
Anterior	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
Labial	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Continuant	-	-	-	+	+	+	+	-	-	-	+	+	+	-	+	+	-	-	+	+	-	-	-	+	-	+
Lateral	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
Nasal	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
Sonorant	-	-	+	-	-	-	-	-	-	+	-	-	+	+	-	-	-	-	+	+	-	-	+	+	-	-
Strident	-	-	-	+	+	-	-	-	-	-	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-
Voiced	-	+	+	-	+	-	+	-	+	+	-	+	+	+	-	+	-	+	+	+	-	+	+	+	-	-

Table 2. Distinctive Features of American English Vowels

i	ɪ	e	ɛ	æ	u	ʊ	o	ɔ	a	ʌ	ə	
+	+	-	-	-	+	+	-	-	-	-	-	**high**
-	-	-	-	+	-	-	-	-	+	+	-	**low**
-	-	-	-	-	+	+	+	+	+	-	-	**back**
-	-	-	-	-	+	+	+	+	-	-	-	**rounded**
+	-	+	-	-	+	-	+	-	-	-	-	**ATR**

Phonemes and phones

Phoneme

Smallest contrastive unit in language.

Abstract idea.

Phone

Instances of phonemes in actual utterances.

Physical segments.

Example: “pat” vs. “bat”

4 phonemes 6 phones

Exploring How Deep Neural Networks Form Phonemic Categories

Tasha Nagamine[1], *Michael L. Seltzer*[2], *Nima Mesgarani*[1]

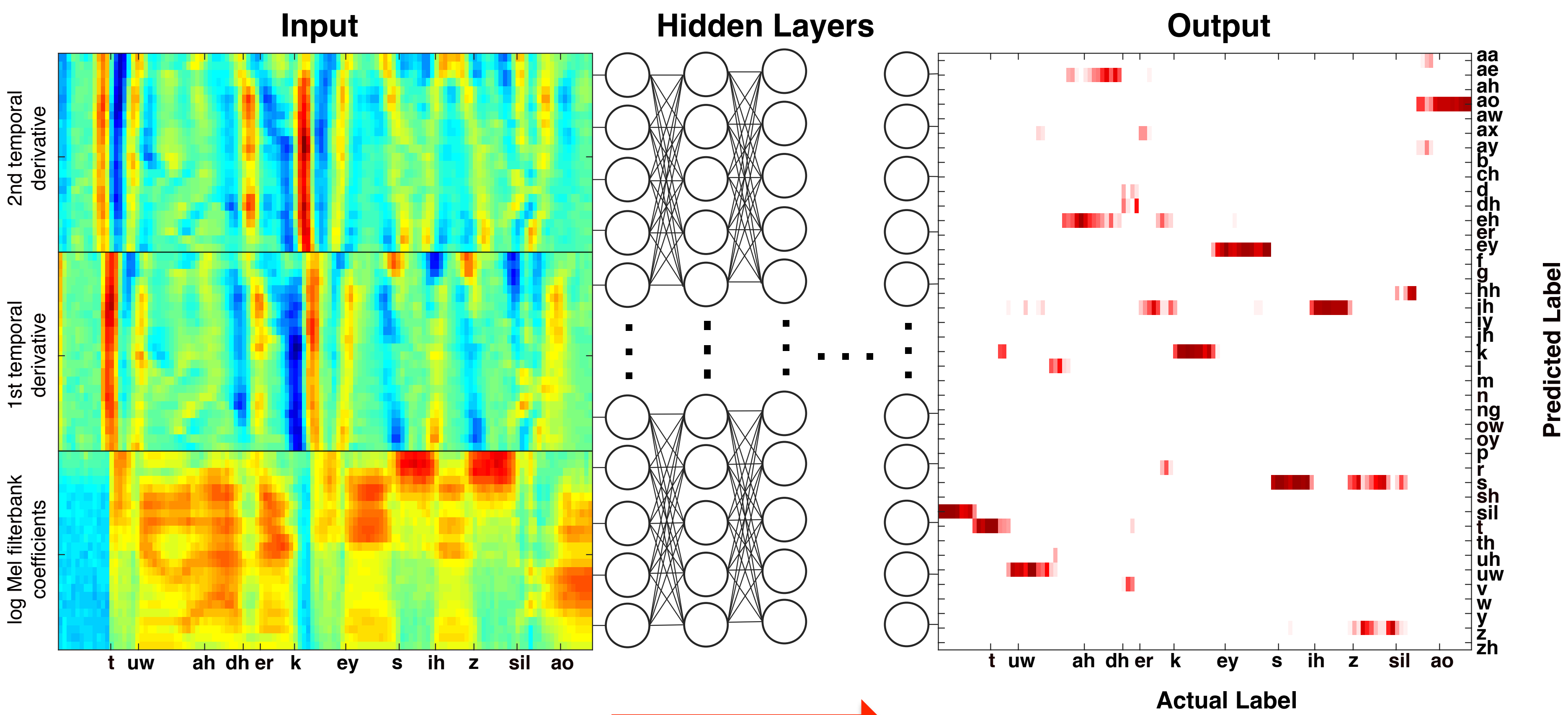
[1]Department of Electrical Engeering, Columbia University, New York, USA
[2]Microsoft Research, Redmond, USA

tasha.nagamine@columbia.edu, mseltzer@microsoft.com, nima@ee.columbia.edu

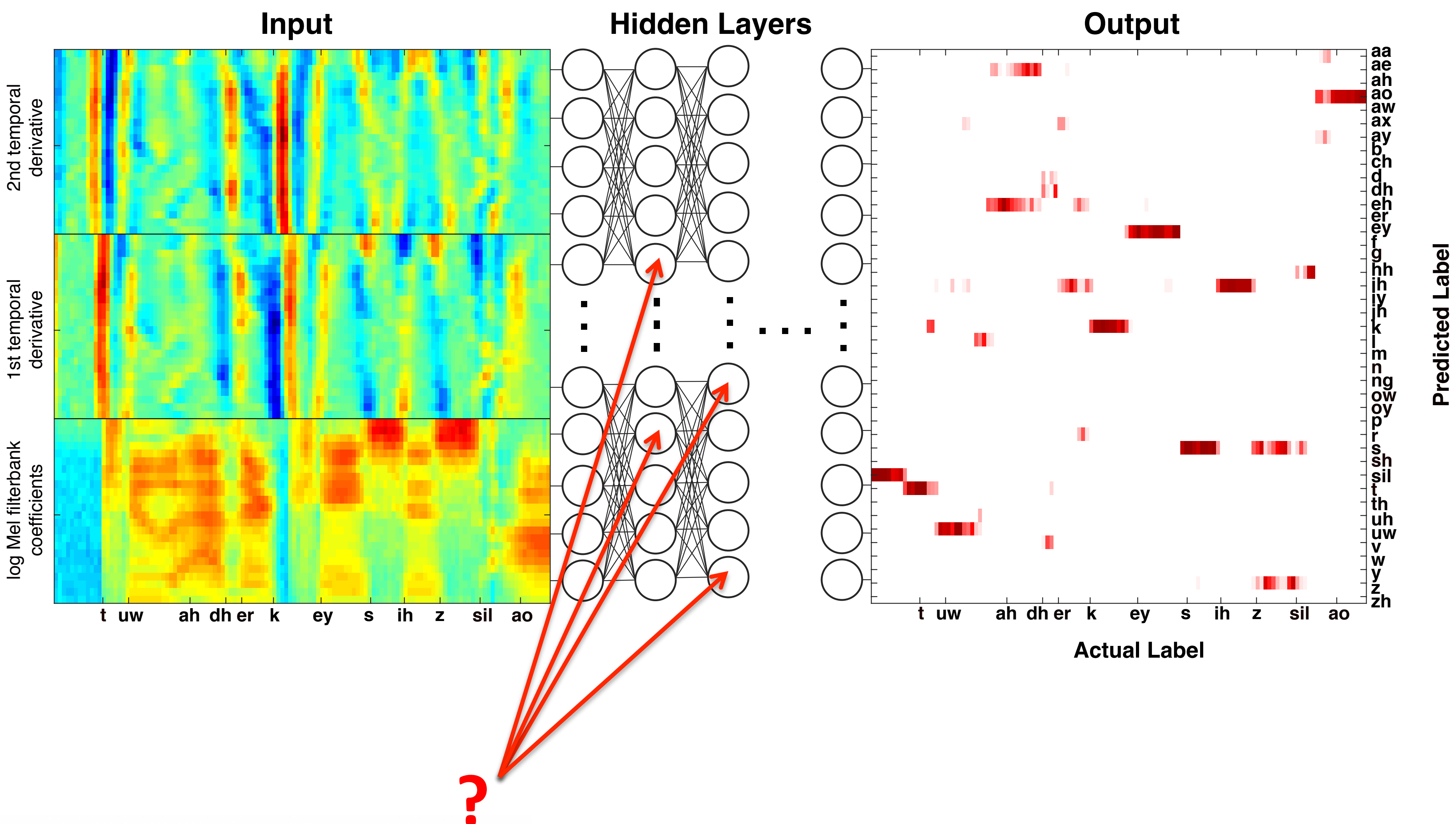
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Microsoft Research

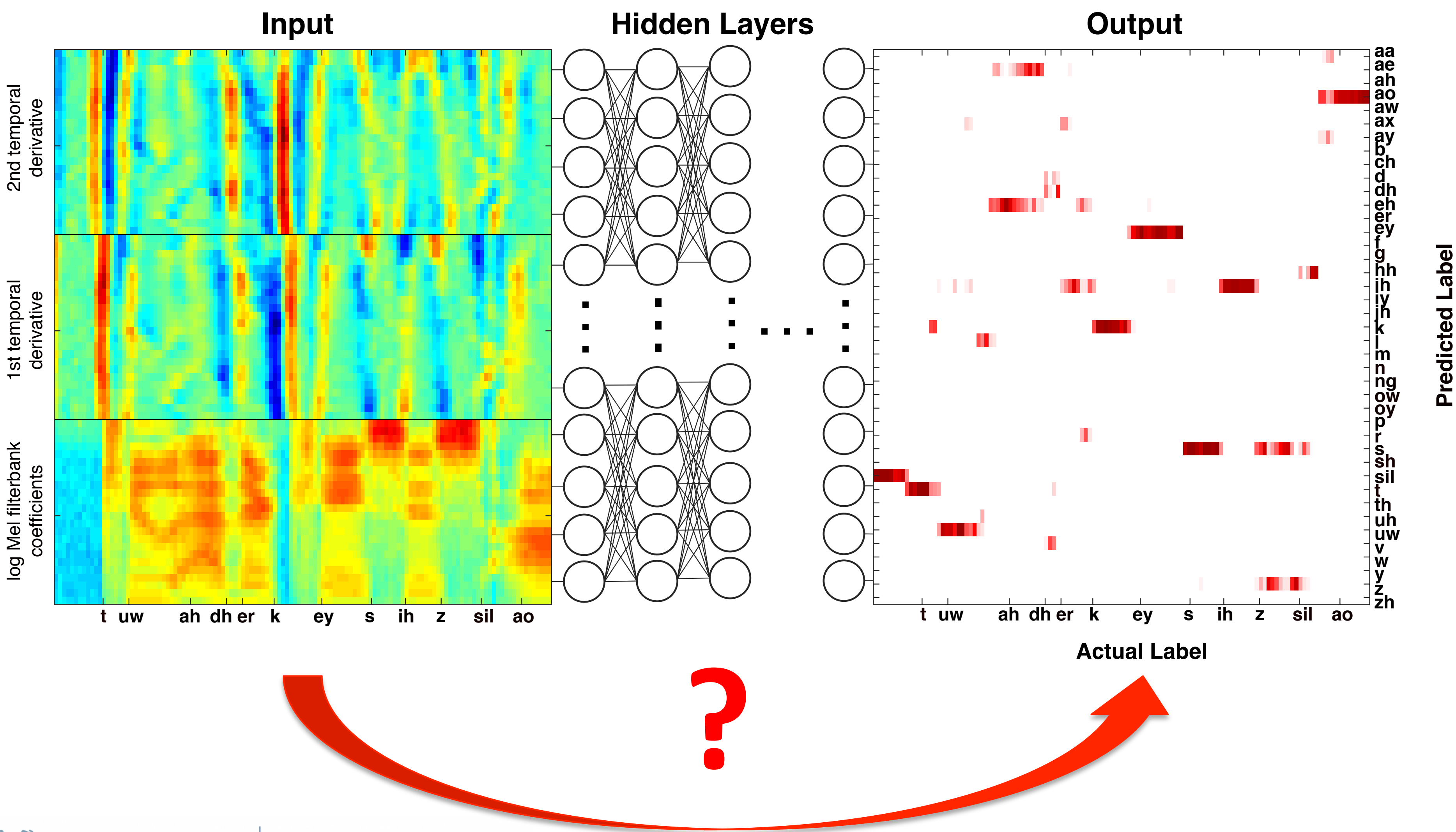
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science



Feed-forward series of nonlinear transformations…



Nagamine et al. Interspeech 2015
October 14, 2015

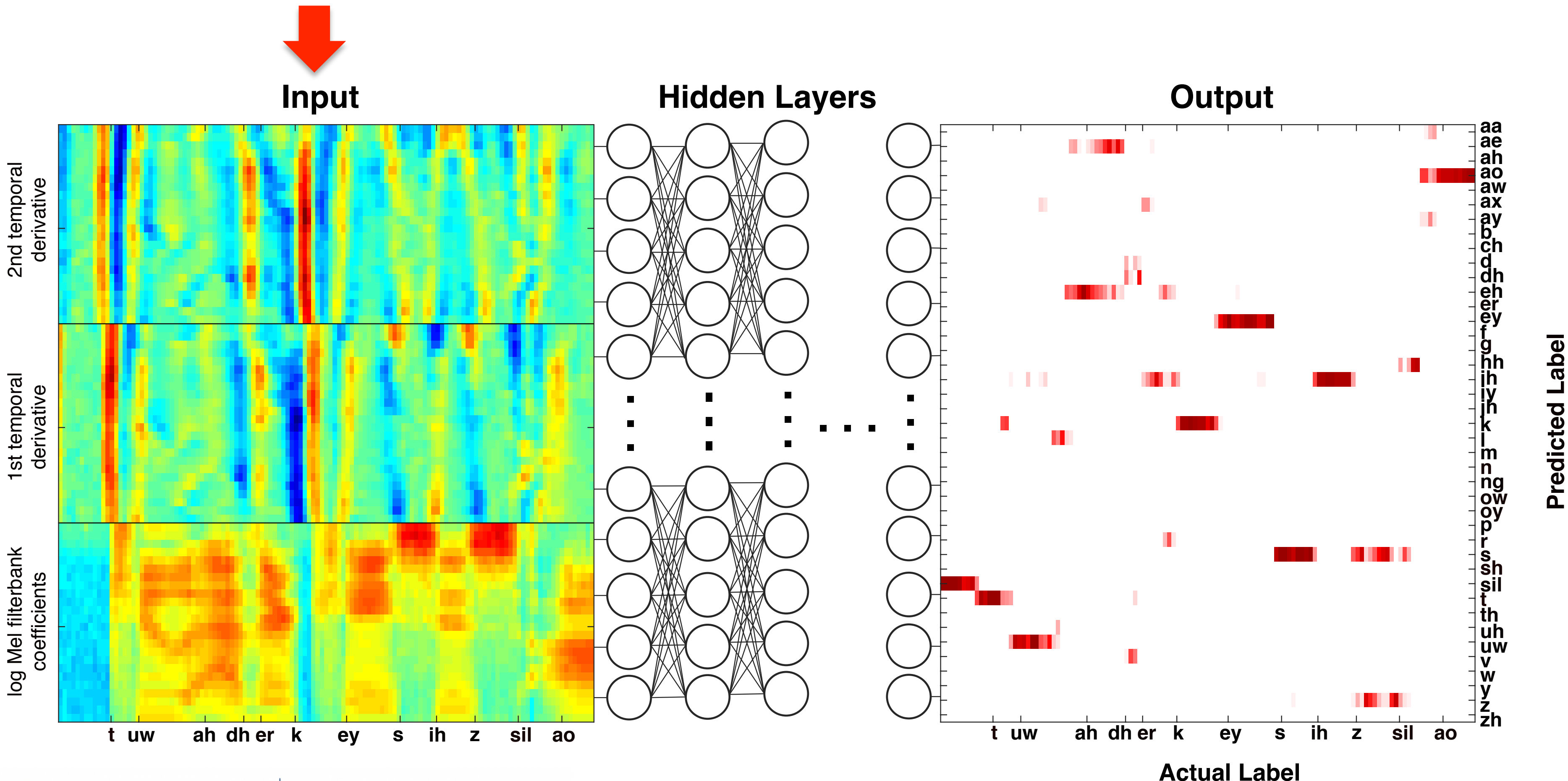


Input
Hidden Layers
Output
2nd temporal derivative
1st temporal derivative
log Mel filterbank coefficients
t uw ah dh er k ey s ih z sil ao
aa ae ah ao aw ax ay b ch d dh eh er ey f g hh ih iy jh k l m n ng ow oy p r s sh sil t th uh uw v w y z zh
Predicted Label
t uw ah dh er k ey s ih z sil ao
Actual Label
?

DNN Architecture

Input layer

11 frames of 24-dimensional log Mel filter bank coefficients + deltas



Nagamine et al. Interspeech 2015

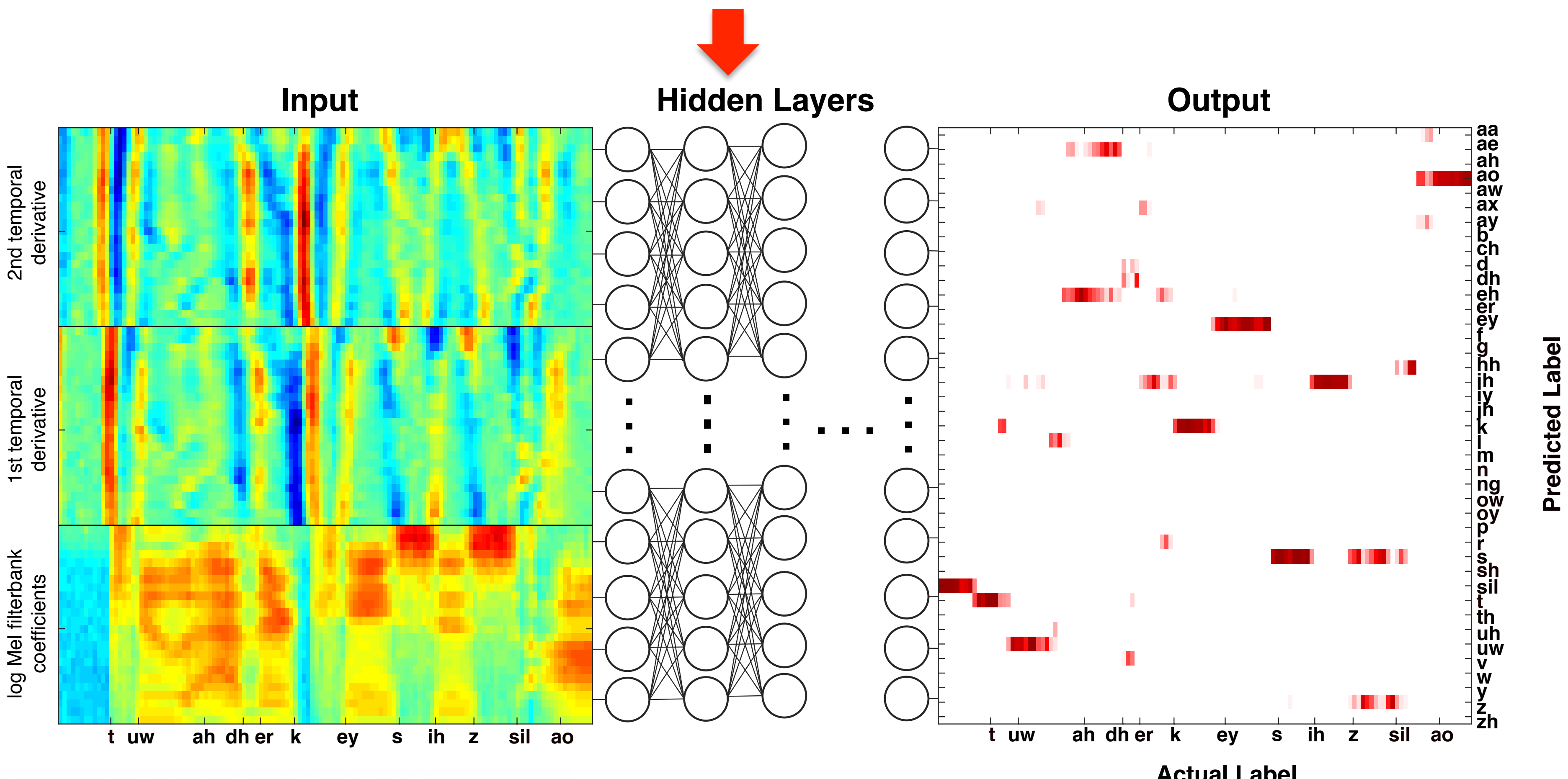
DNN Architecture

Input layer

11 frames of 24-dimensional log Mel filter bank coefficients + deltas

5 sigmoid hidden layers

256 nodes each; fully connected feed-forward



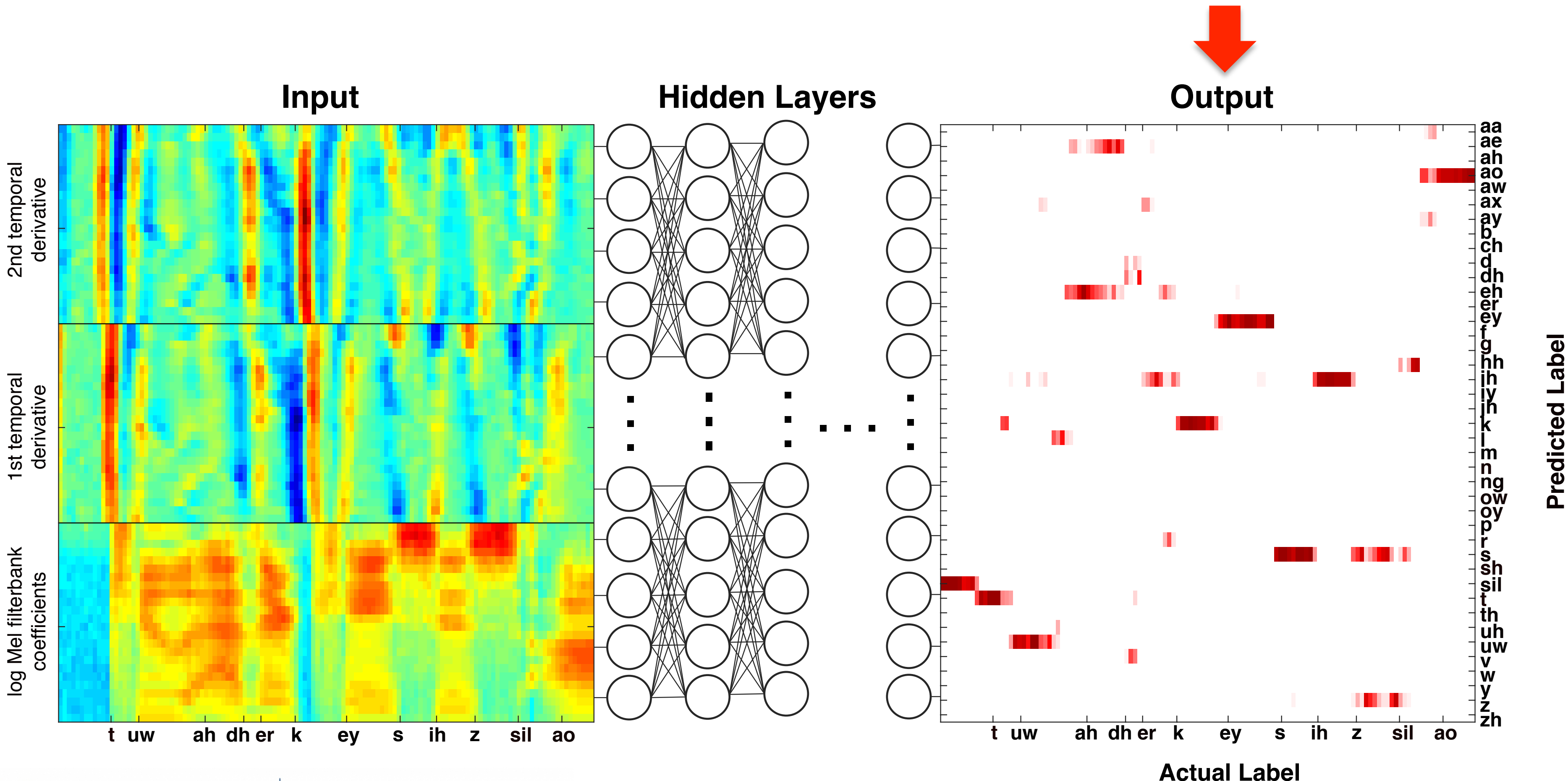
Nagamine et al. Interspeech 2015

DNN Architecture

Input layer
11 frames of 24-dimensional log Mel filter bank coefficients + deltas

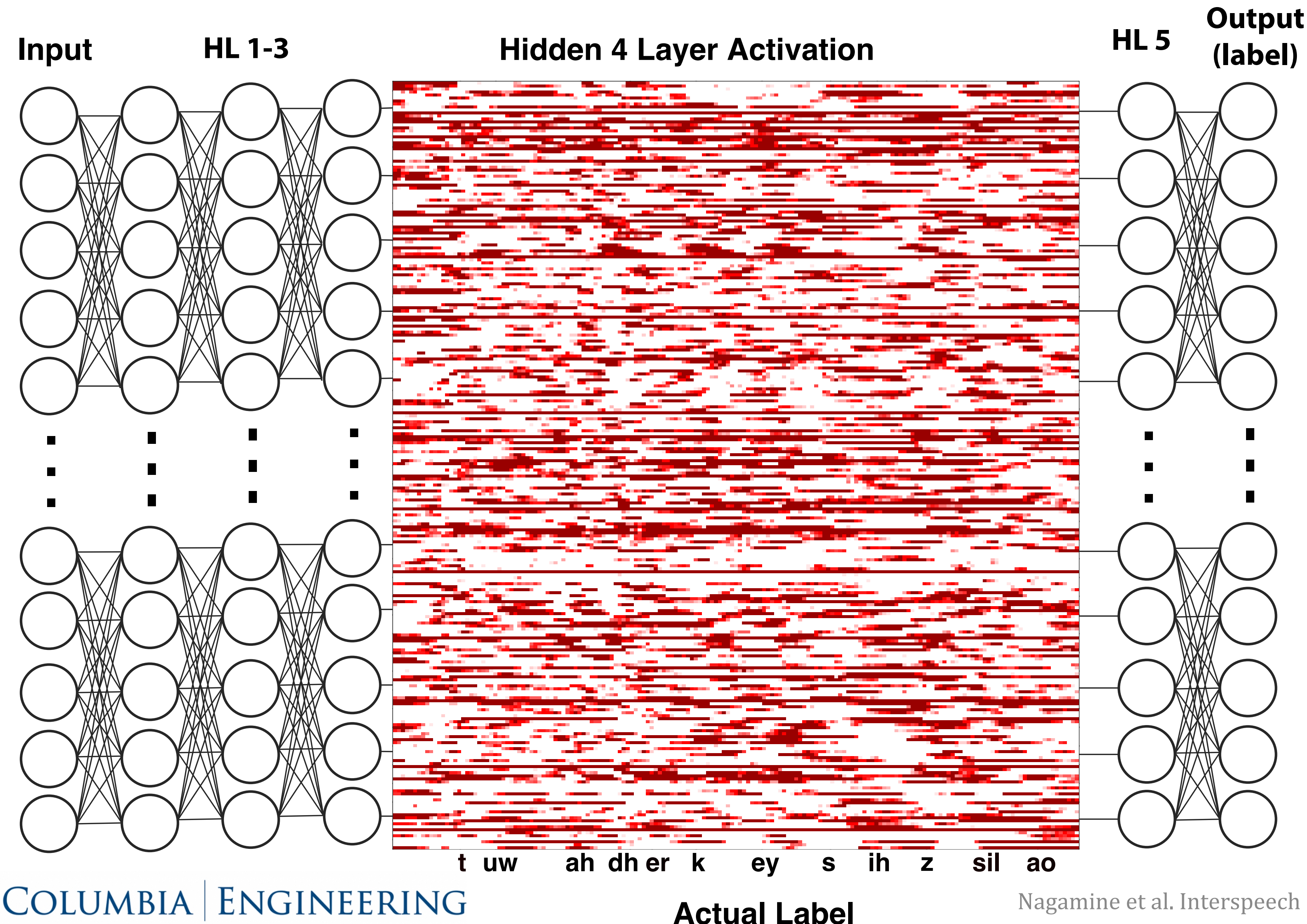
5 sigmoid hidden layers
256 nodes each; fully connected feed-forward

Softmax output layer
41 nodes for 40 phonemes and silence; context independent

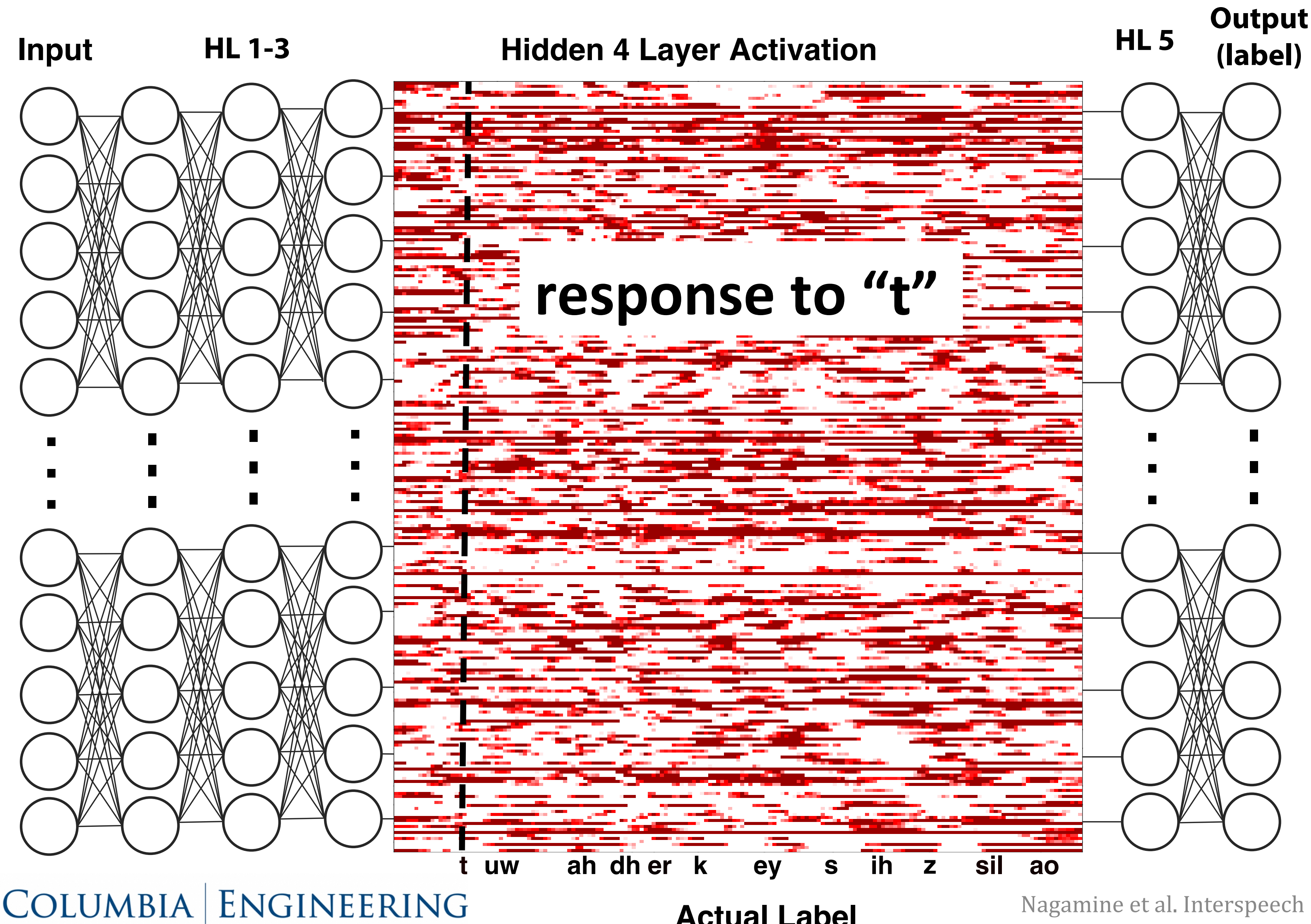


Nagamine et al. Interspeech 2015

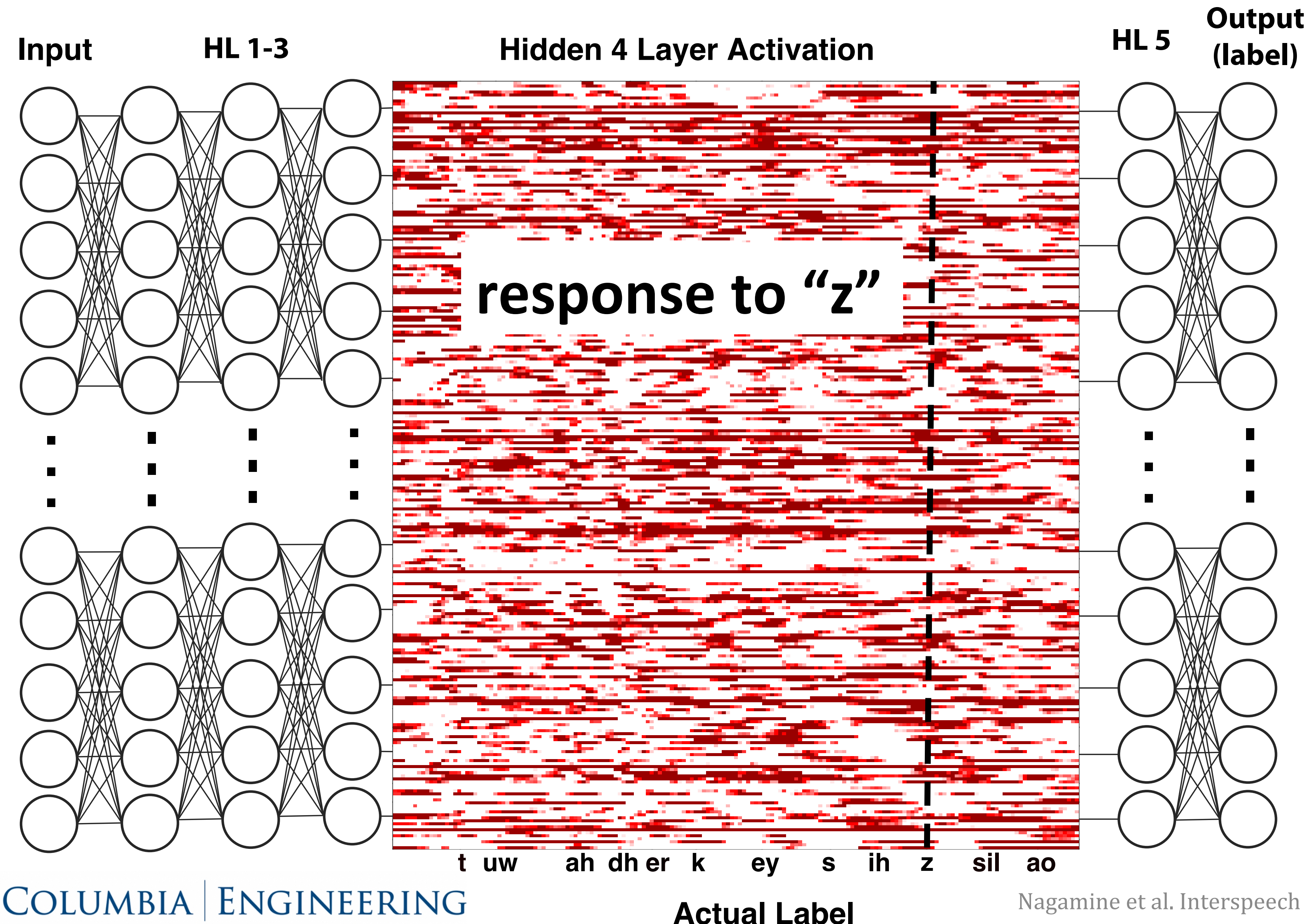
Speech stimuli & DNN activations



Speech stimuli & DNN activations

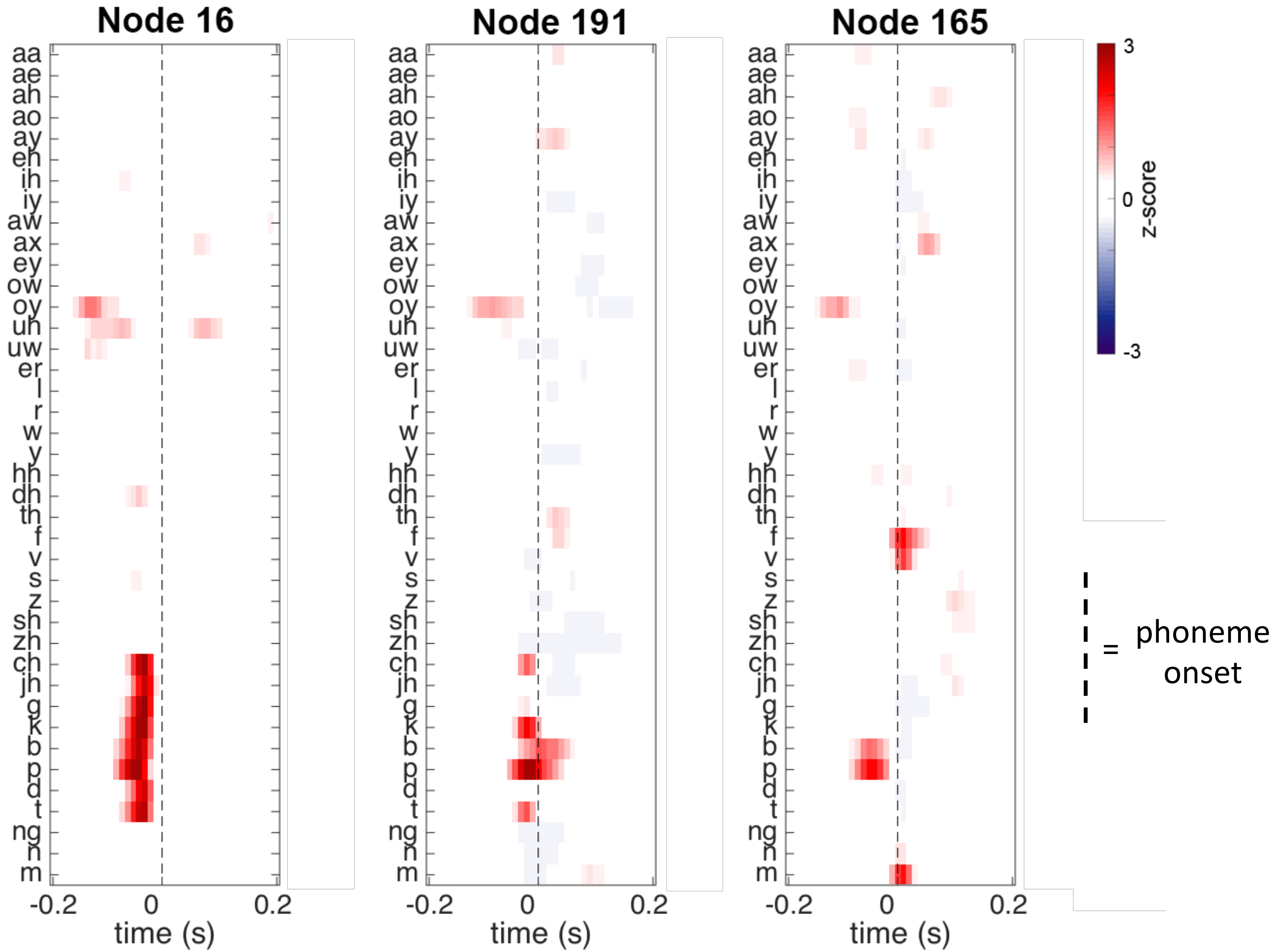


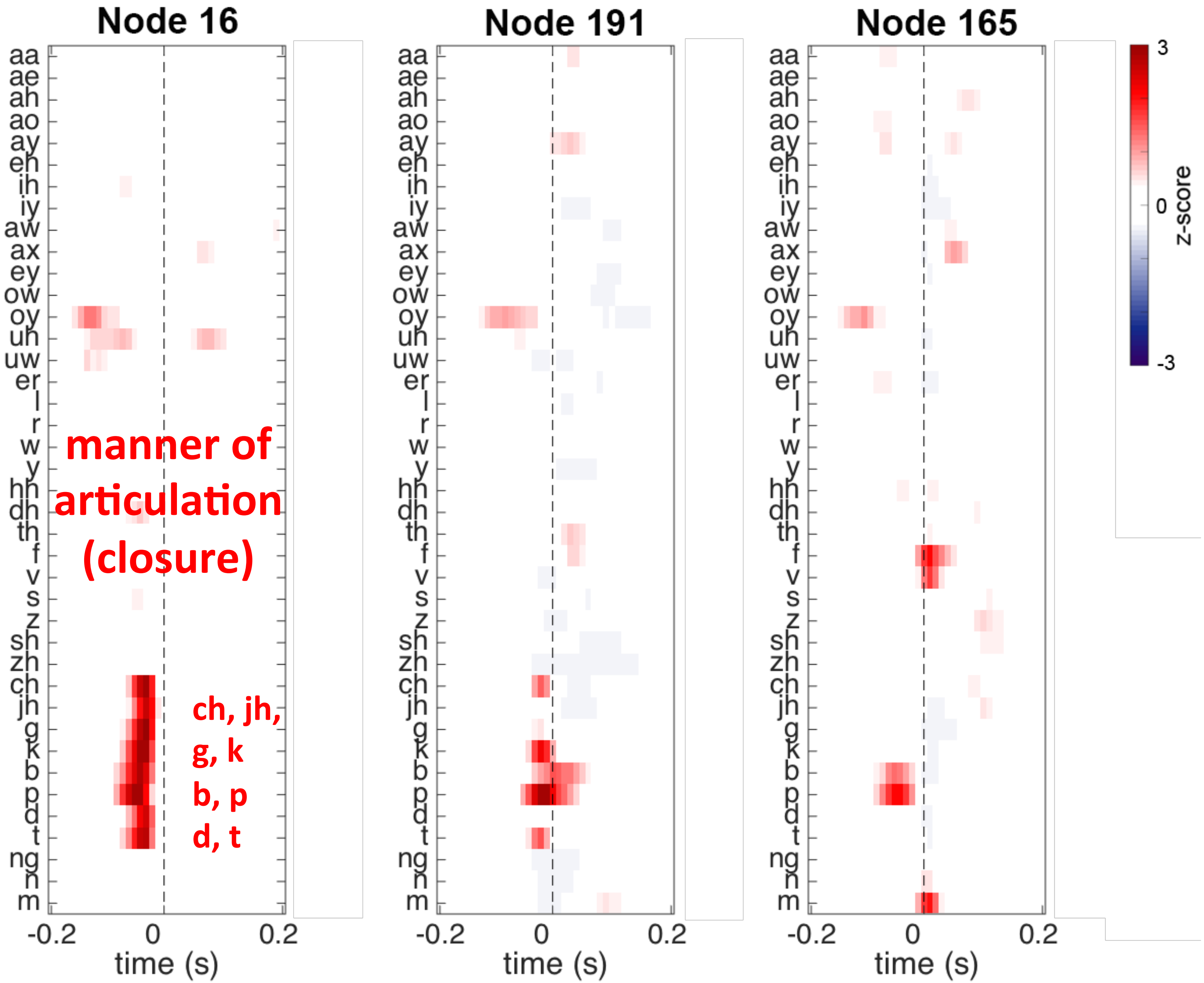
Speech stimuli & DNN activations



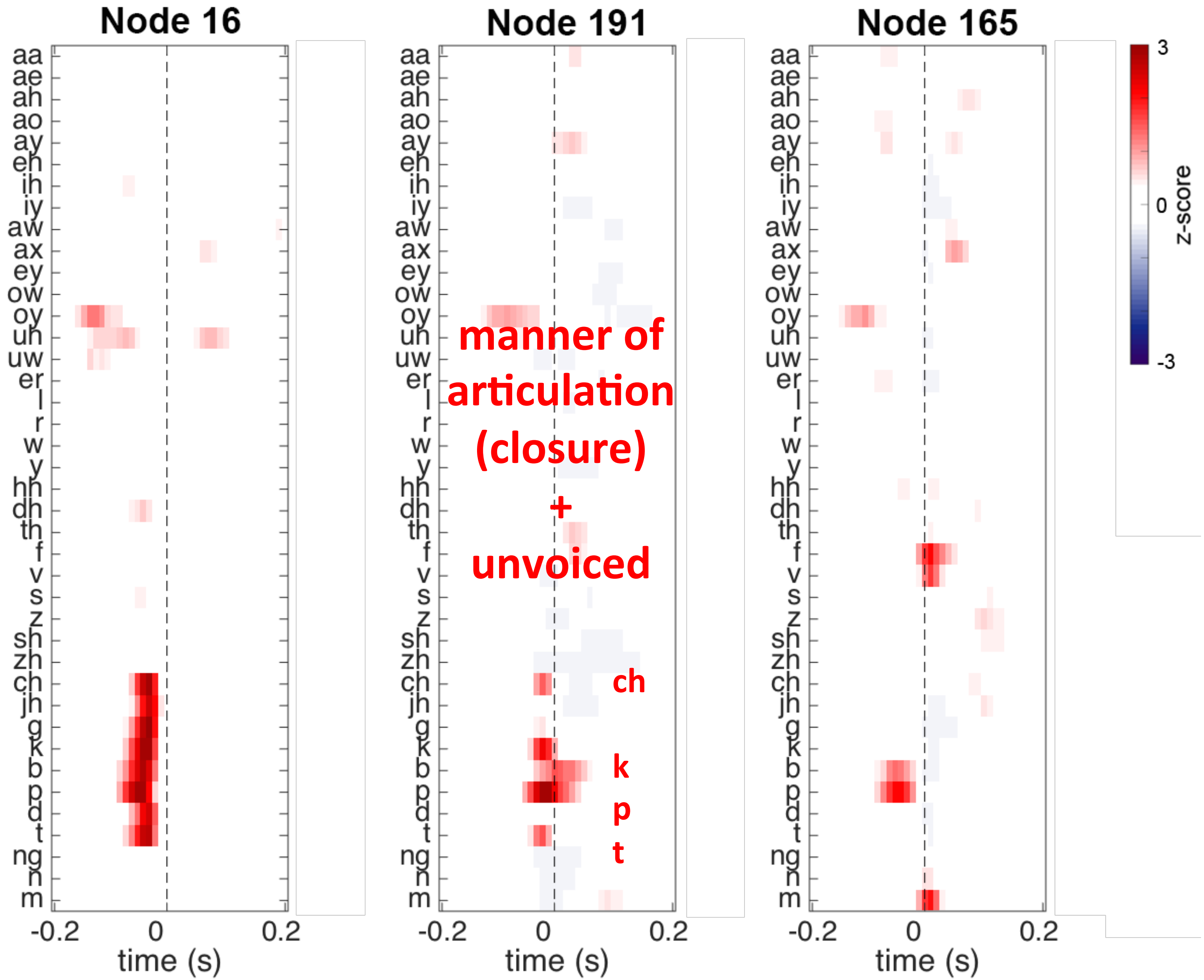
Summary of findings

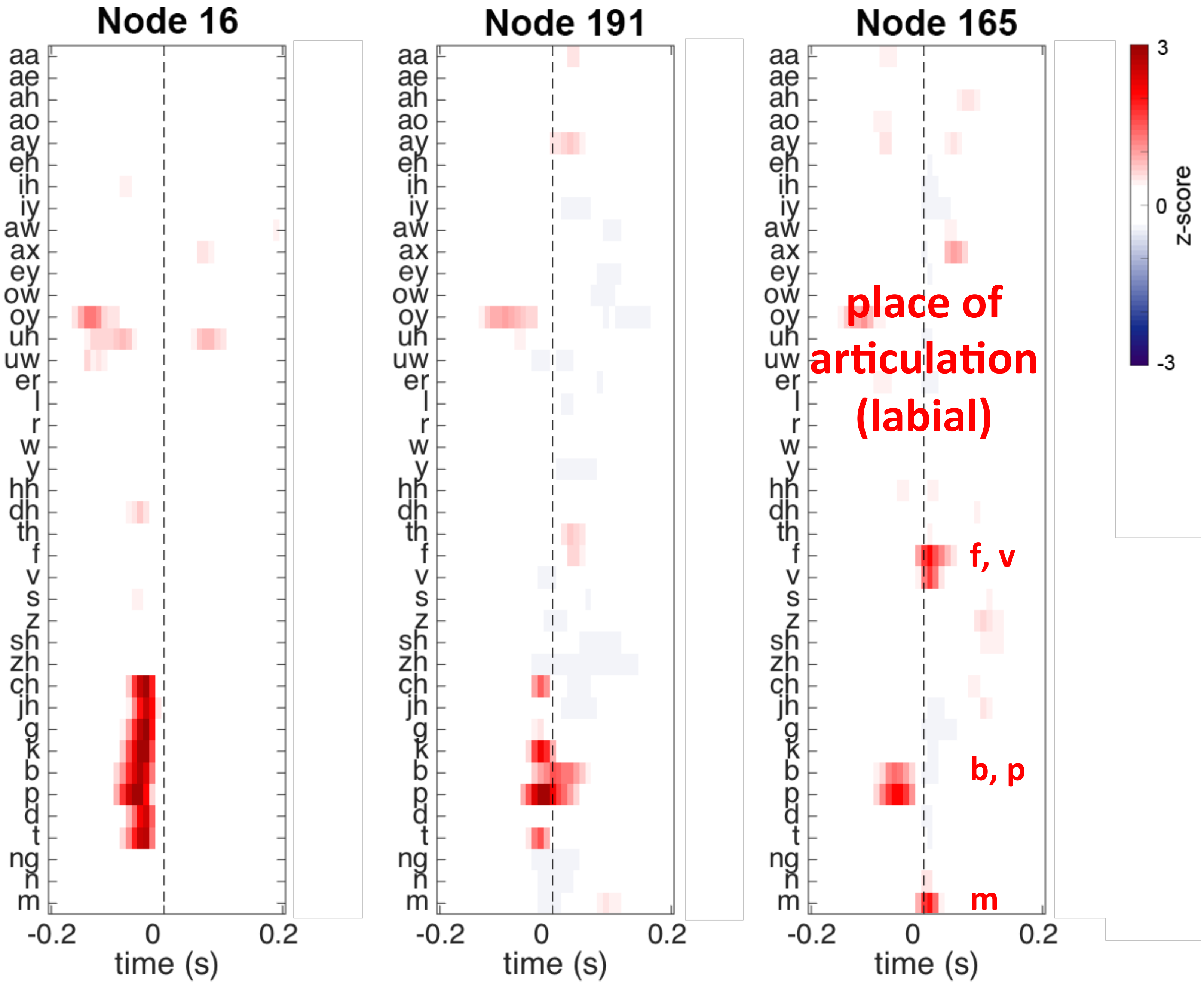
1. Nodes are selective to phonetic features at the individual and population level


Node 16
Node 191
Node 165
aa
ae
ah
ao
ay
eh
ih
iy
aw
ax
ey
ow
oy
uh
uw
er
l
r
w
y
hh
dh
th
f
v
s
z
sh
zh
ch
jh
g
k
b
p
d
t
ng
n
m
-0.2
0
0.2
time (s)
3
0
-3
z-score
= phoneme onset

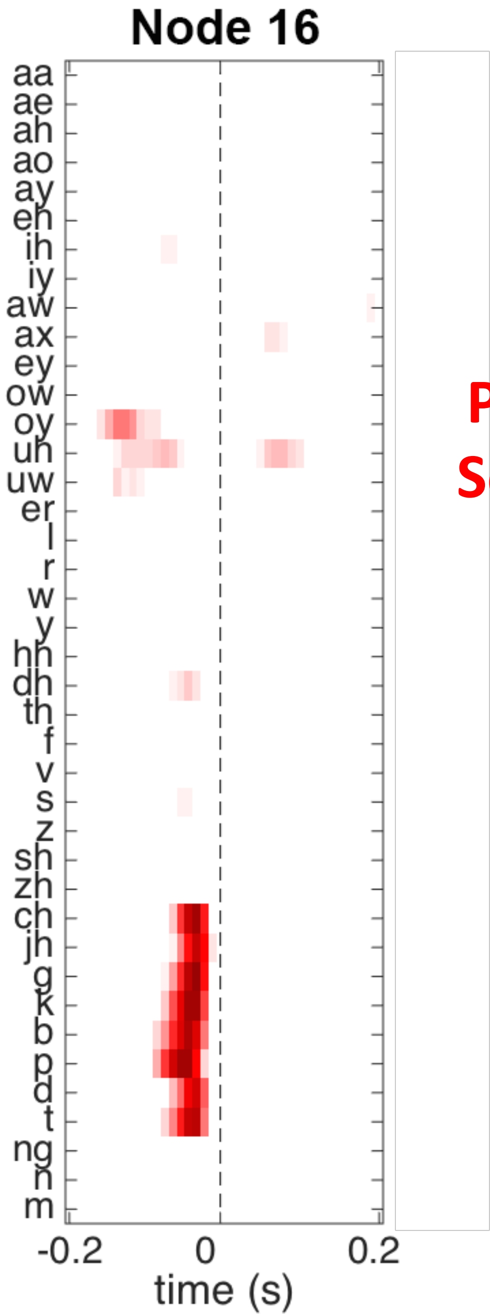


Nagamine et al. Interspeech 2015
October 14, 2015

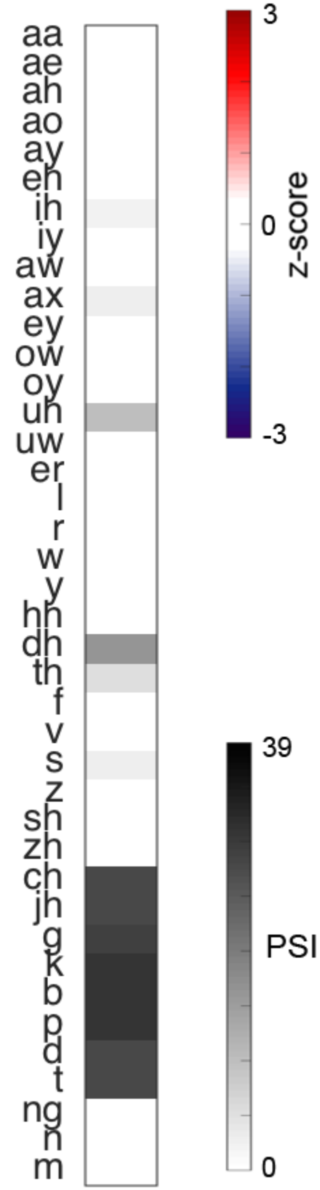

Node 16
Node 191
Node 165
aa
ae
ah
ao
ay
eh
ih
iy
aw
ax
ey
ow
oy
uh
uw
er
l
r
w
y
hh
dh
th
f
v
s
z
sh
zh
ch
jh
g
k
b
p
d
t
ng
n
m
-0.2
0
0.2
time (s)
3
0
-3
z-score
manner of articulation (closure)
+
unvoiced
ch
k
p
t



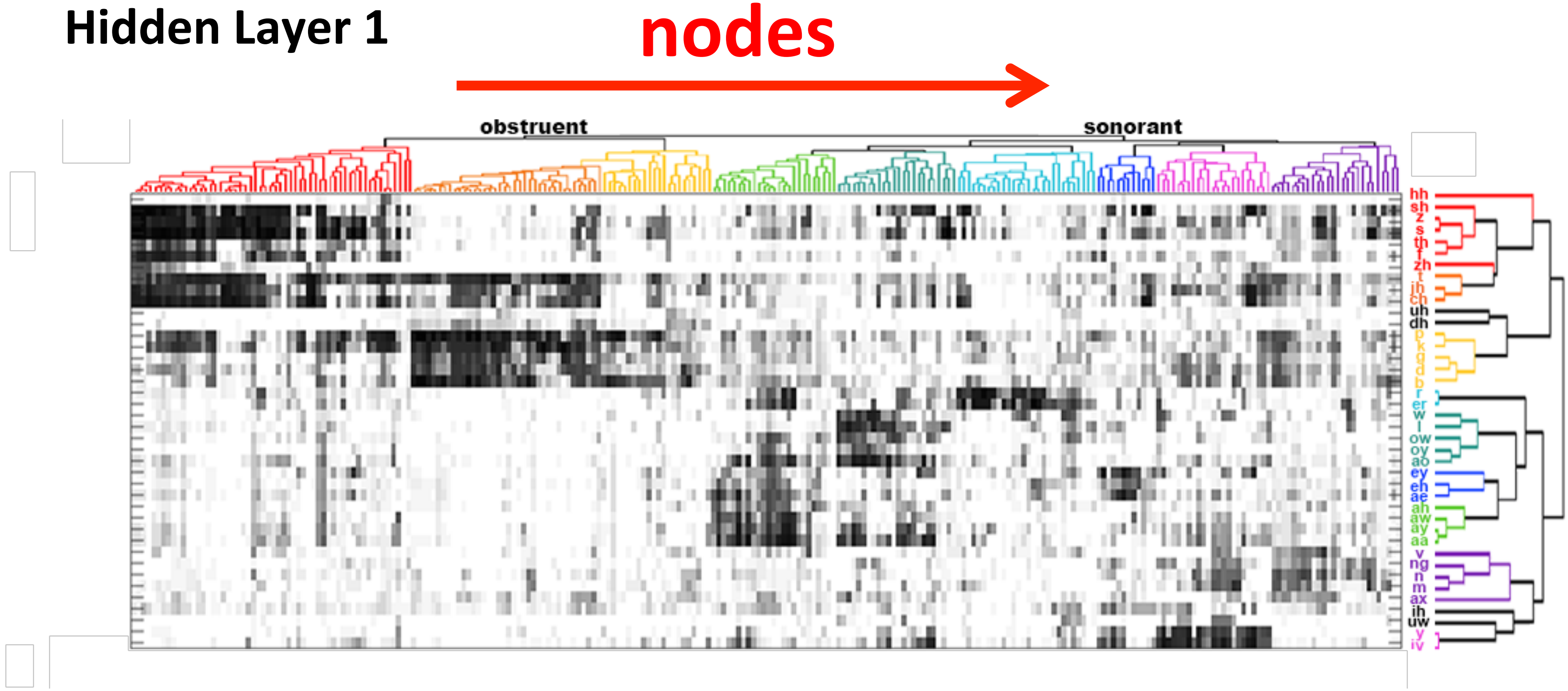
Nagamine et al. Interspeech 2015
October 14, 2015



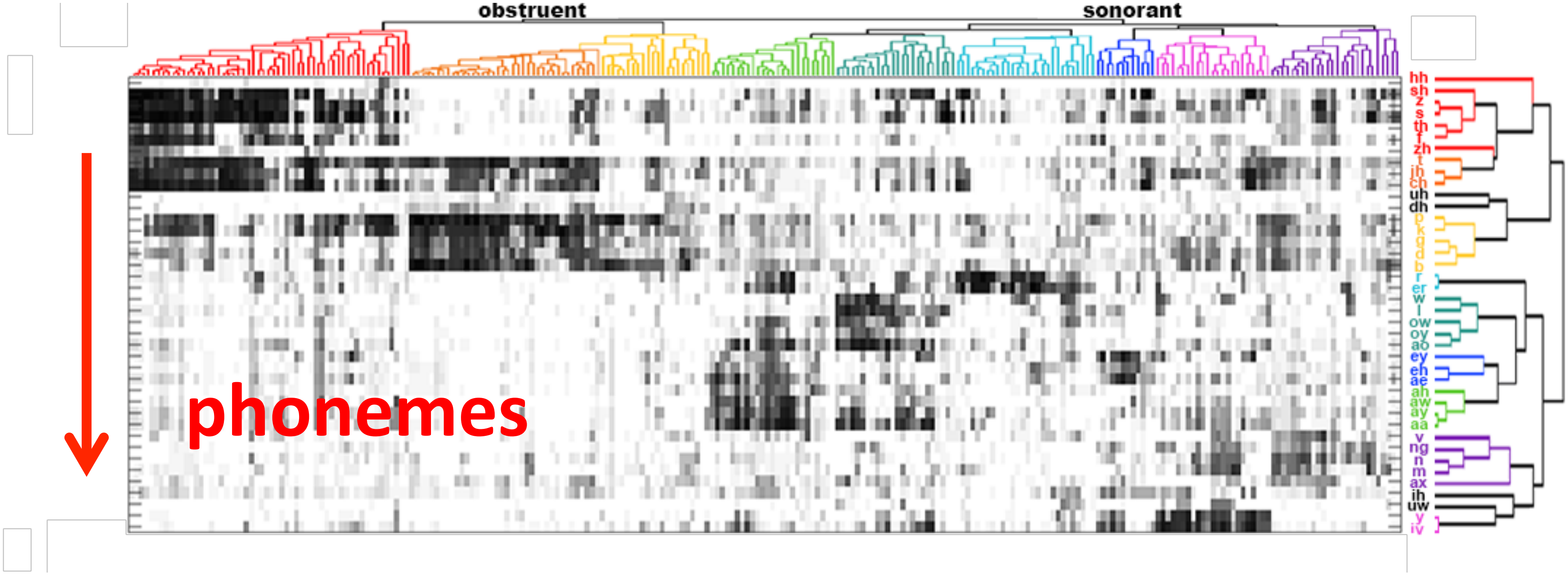
Phoneme Selectivity Index (PSI)



Hidden Layer 1

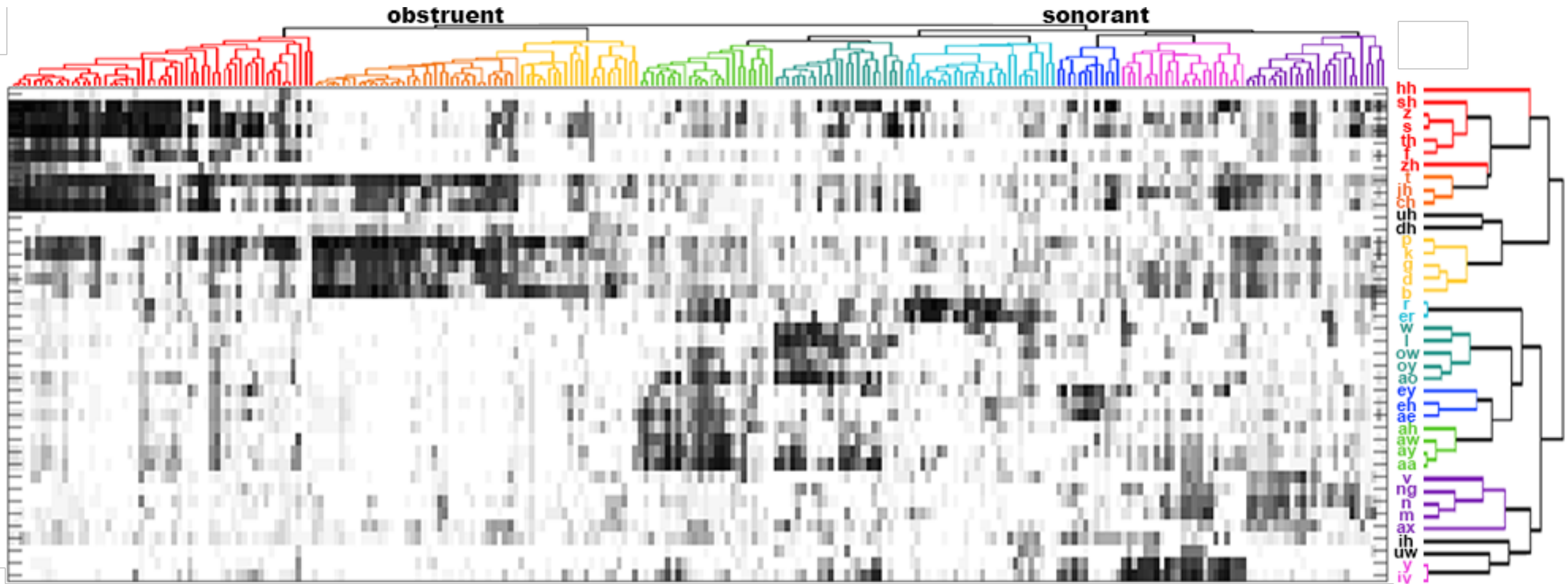


Hidden Layer 1

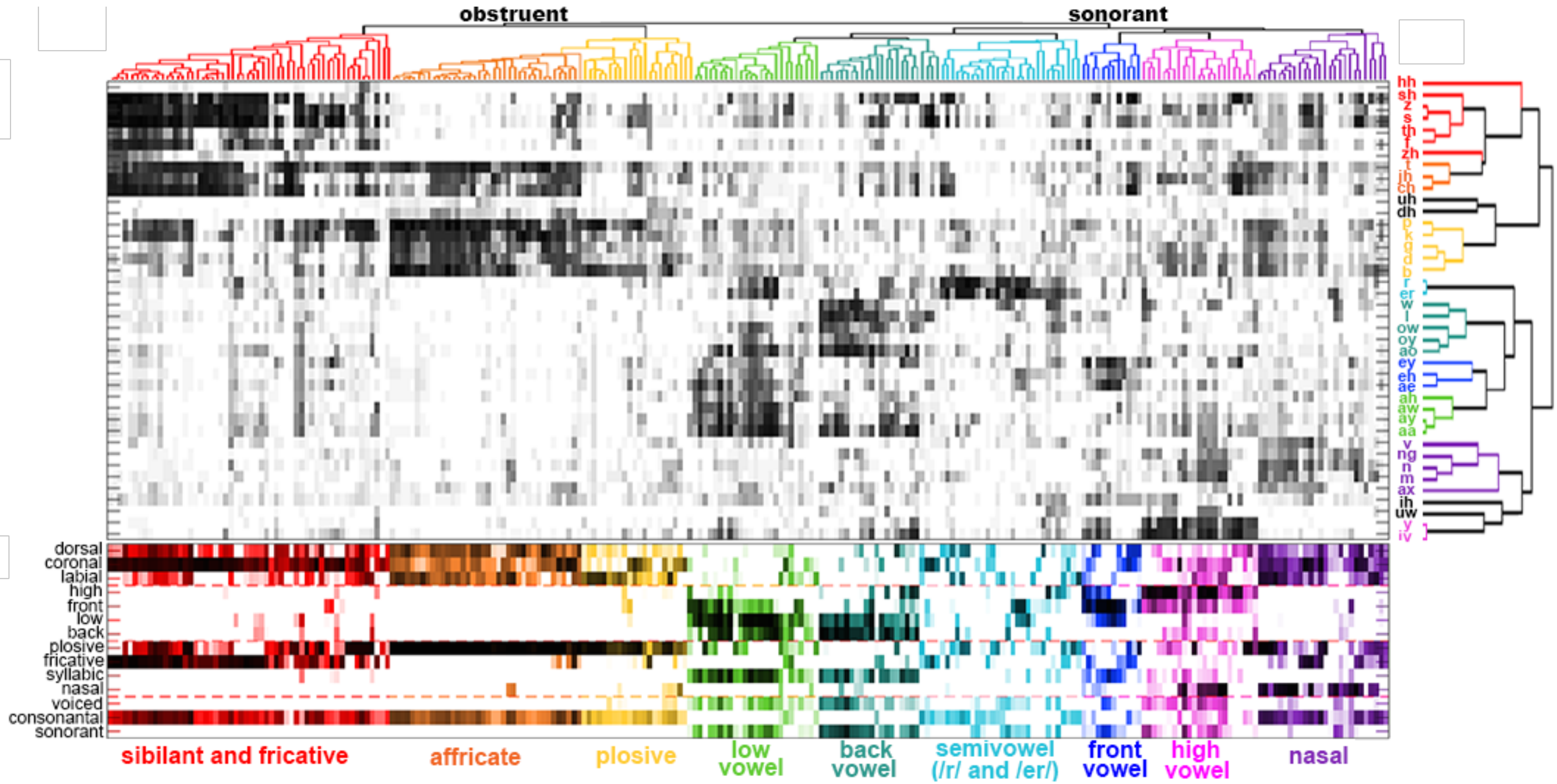


Nagamine et al. Interspeech 2015

Hidden Layer 1

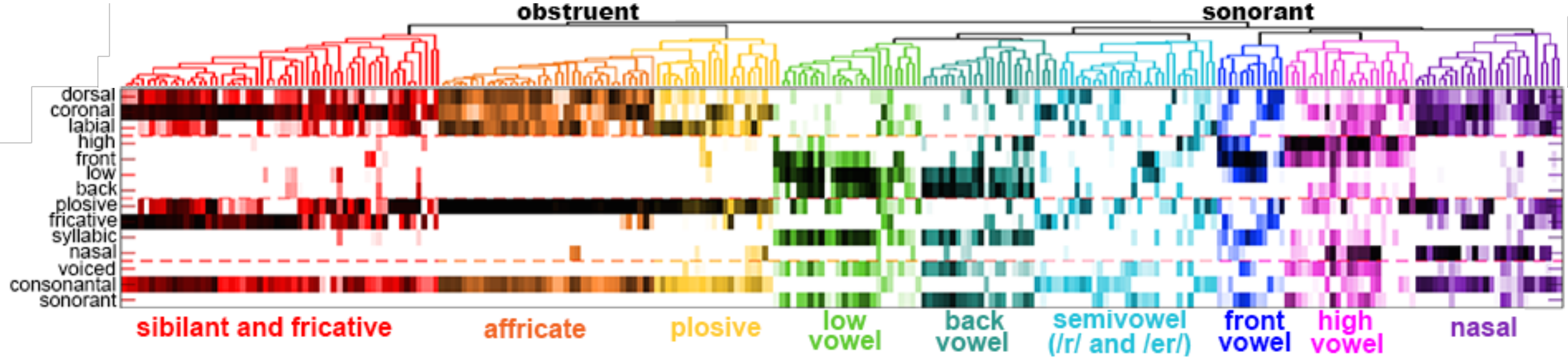


Hidden Layer 1

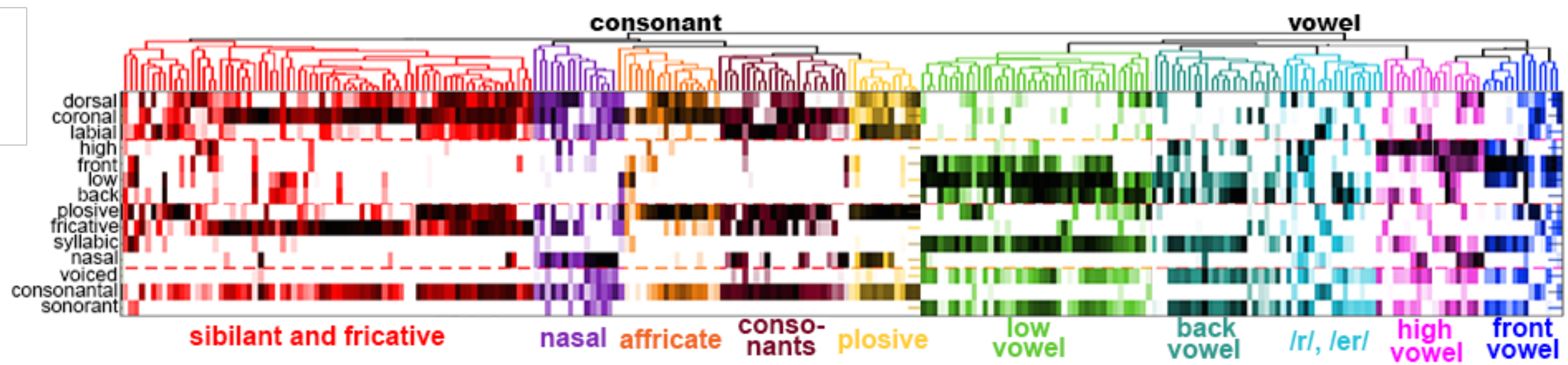


Nagamine et al. Interspeech 2015

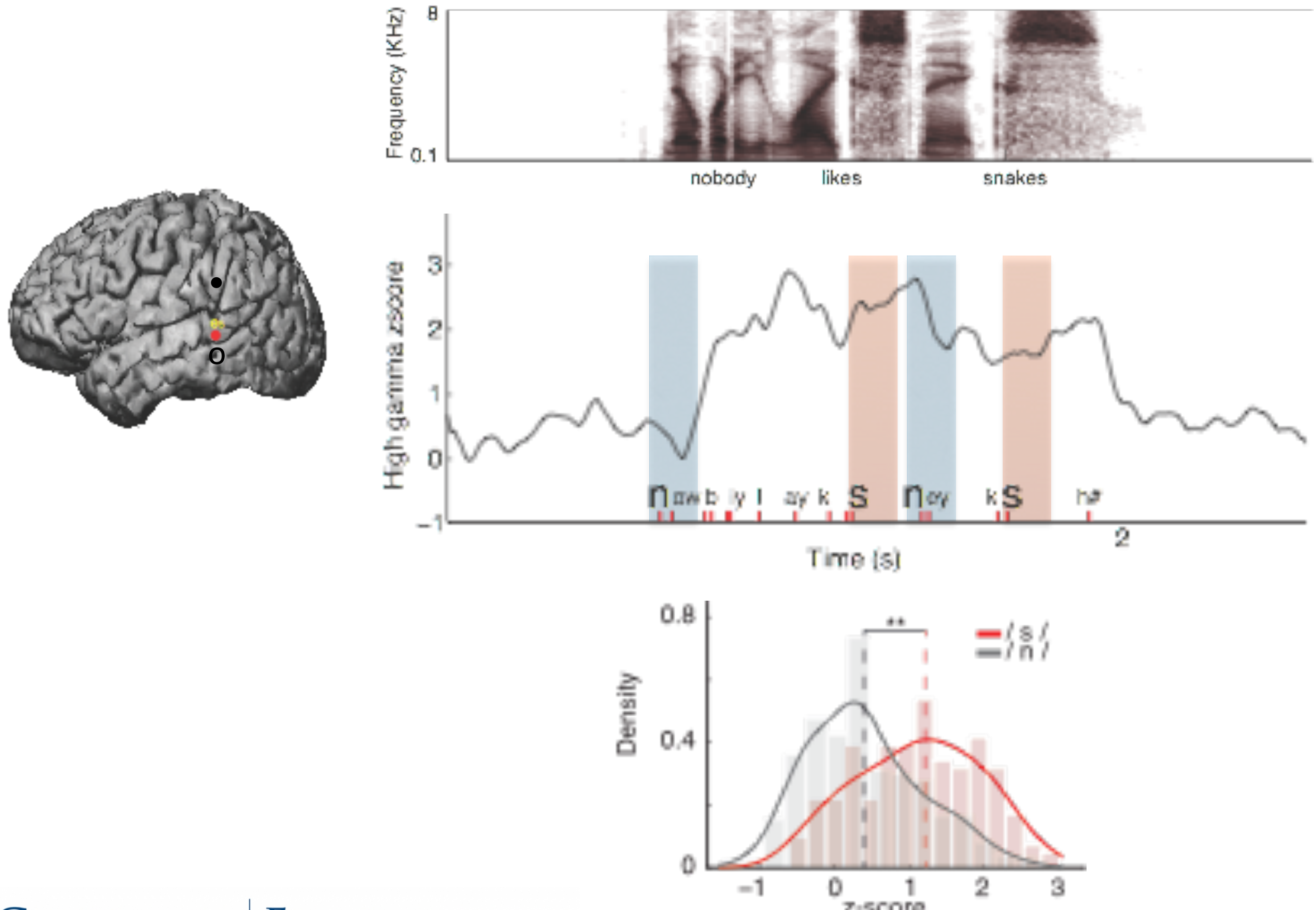
Hidden Layer 1



Hidden Layer 5

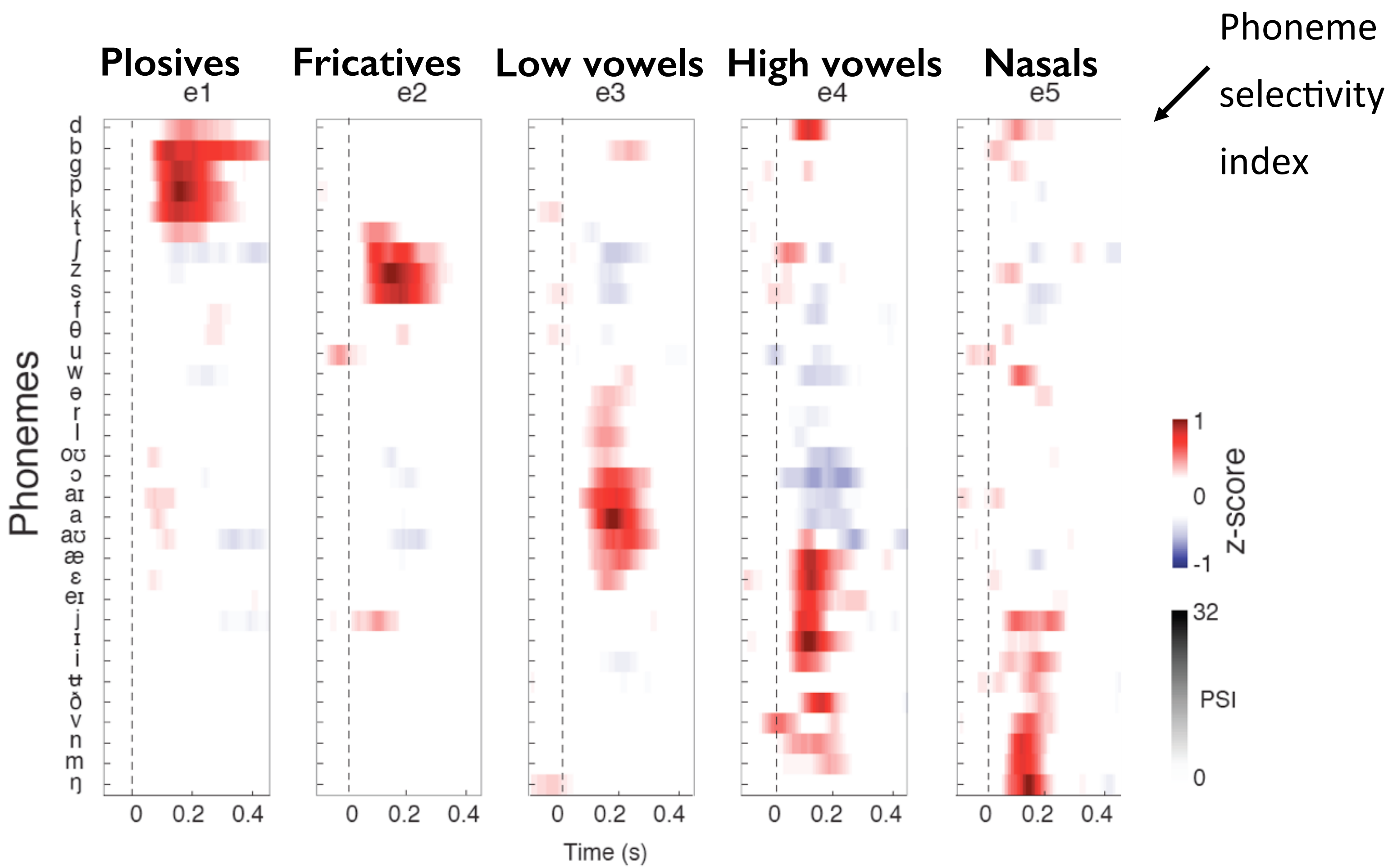


Neural responses to speech in human superior temporal gyrus (STG)



Mesgarani et al., Science 2014

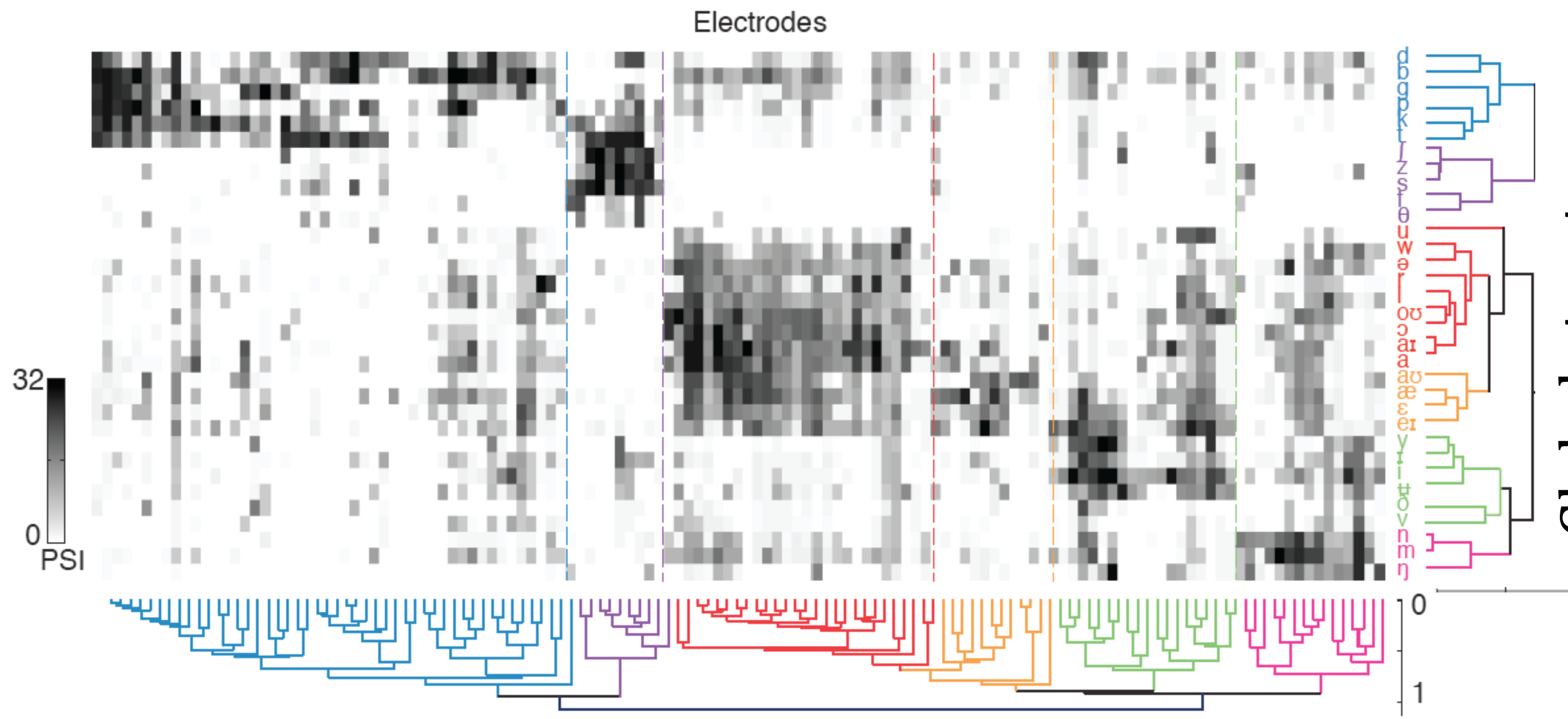
Examples of average phoneme responses in STG



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

Mesgarani et al., Science 2014

Clustering the PSI vectors



Global structures (population)

Local structures (single electrode)

Place

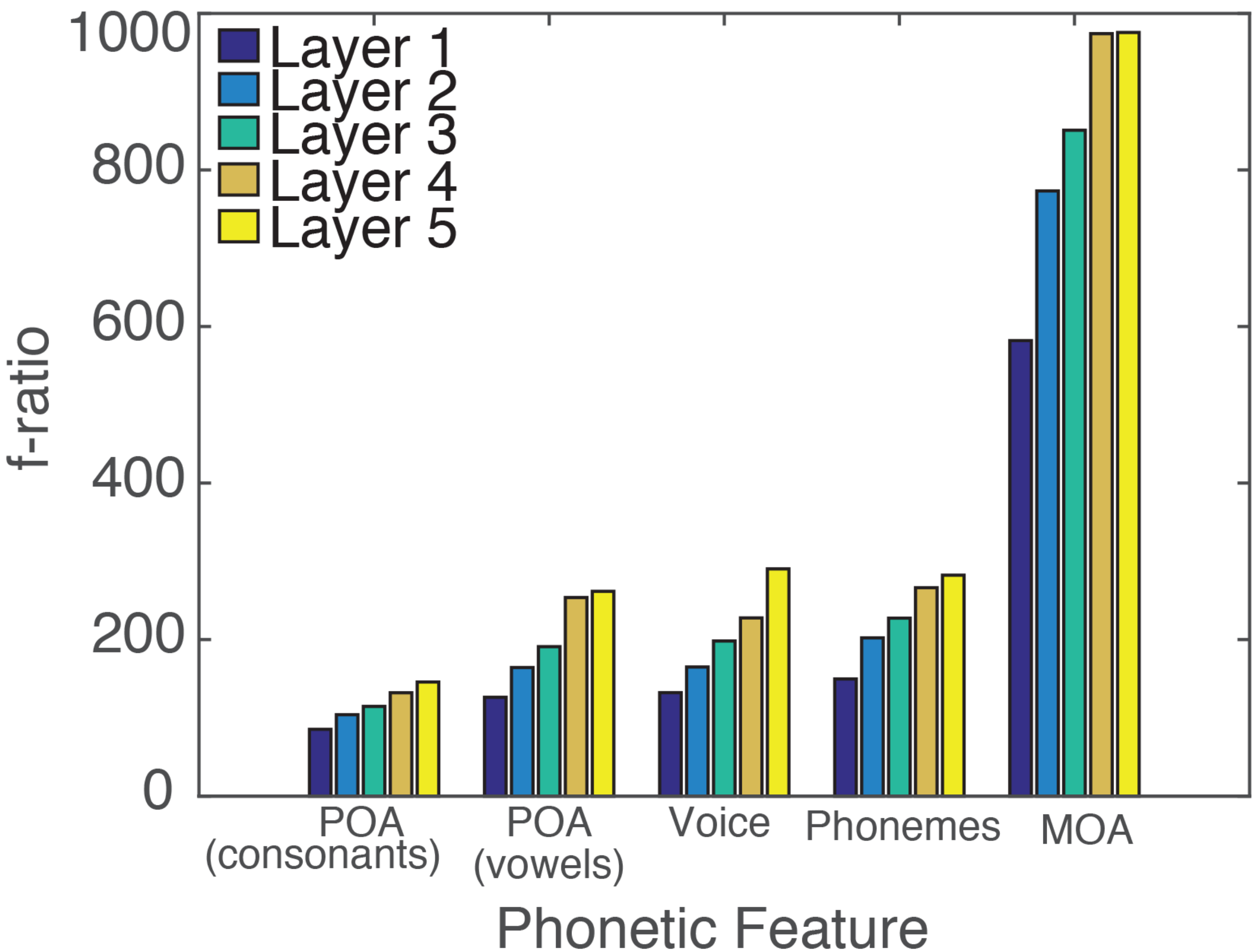
Manner

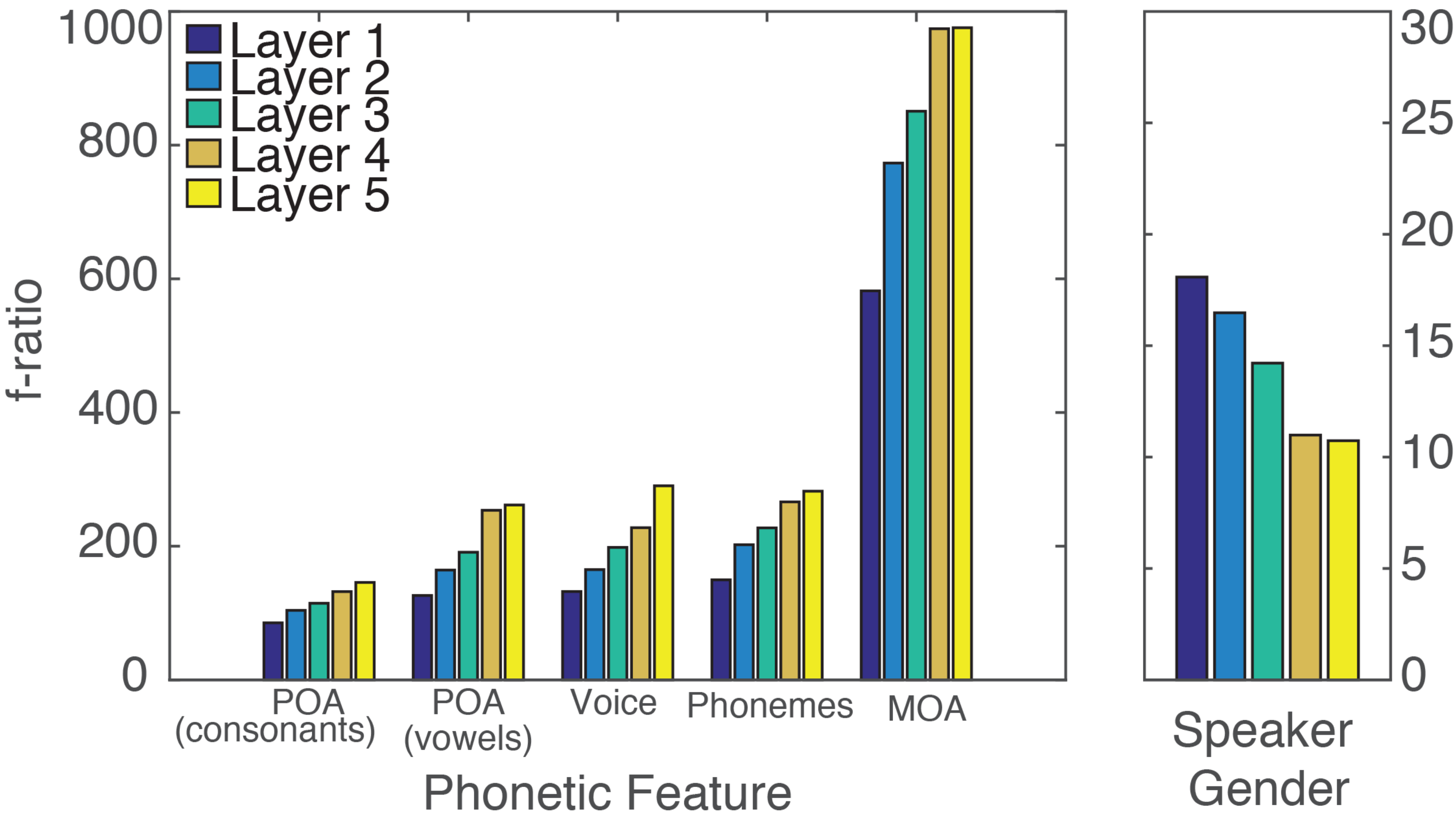
Mesgarani et al., Science 2014

Summary of findings

1. Single nodes and populations of nodes in a layer are selective to phonetic features

2. Phonetic feature encoding becomes more explicit in deeper layers


1000
800
600
400
200
0
f-ratio
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
POA (consonants)
POA (vowels)
Voice
Phonemes
MOA
Phonetic Feature


f-ratio
1000
800
600
400
200
0
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
POA (consonants)
POA (vowels)
Voice
Phonemes
MOA
Phonetic Feature
30
25
20
15
10
5
0
Speaker Gender

Summary of findings

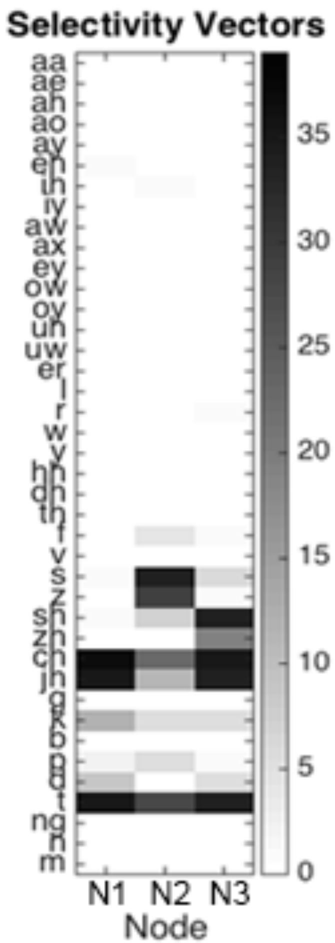
1. Single nodes and populations of nodes in a layer are selective to phonetic features

2. Node selectivity to phonetic features becomes more explicit in deeper layers

3. Network invariance is learned through explicit representation of sources of variability

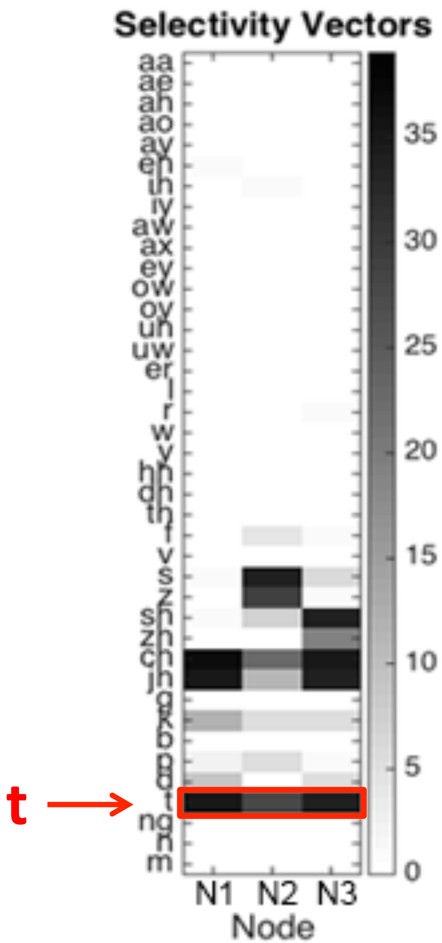
phoneme = “t”

example selectivity for three nodes (N1, N2, N3)



phoneme = “t”

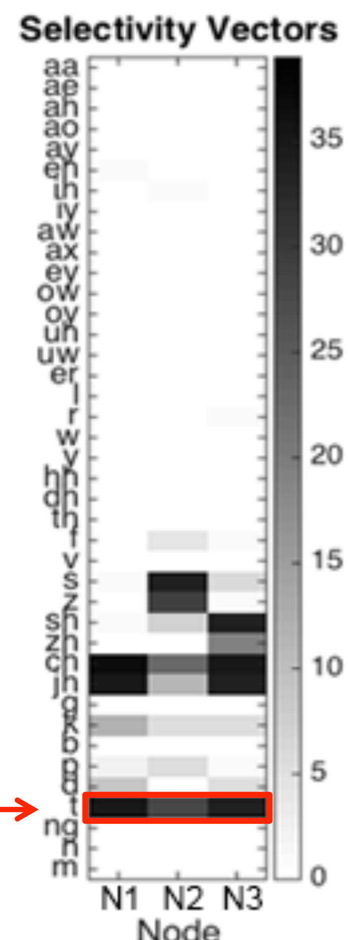
example selectivity for three nodes (N1, N2, N3)



Nagamine et al. Interspeech 2015
October 14, 2015

phoneme = “t”

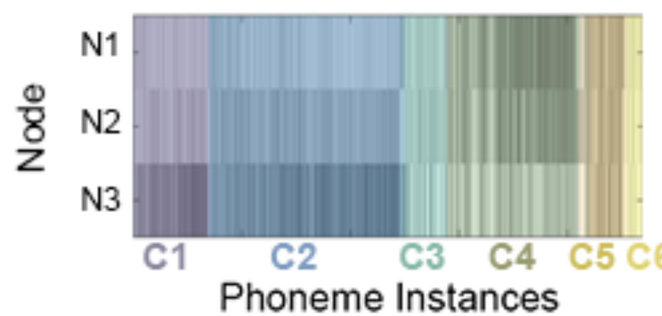
example selectivity for three nodes (N1, N2, N3)



t

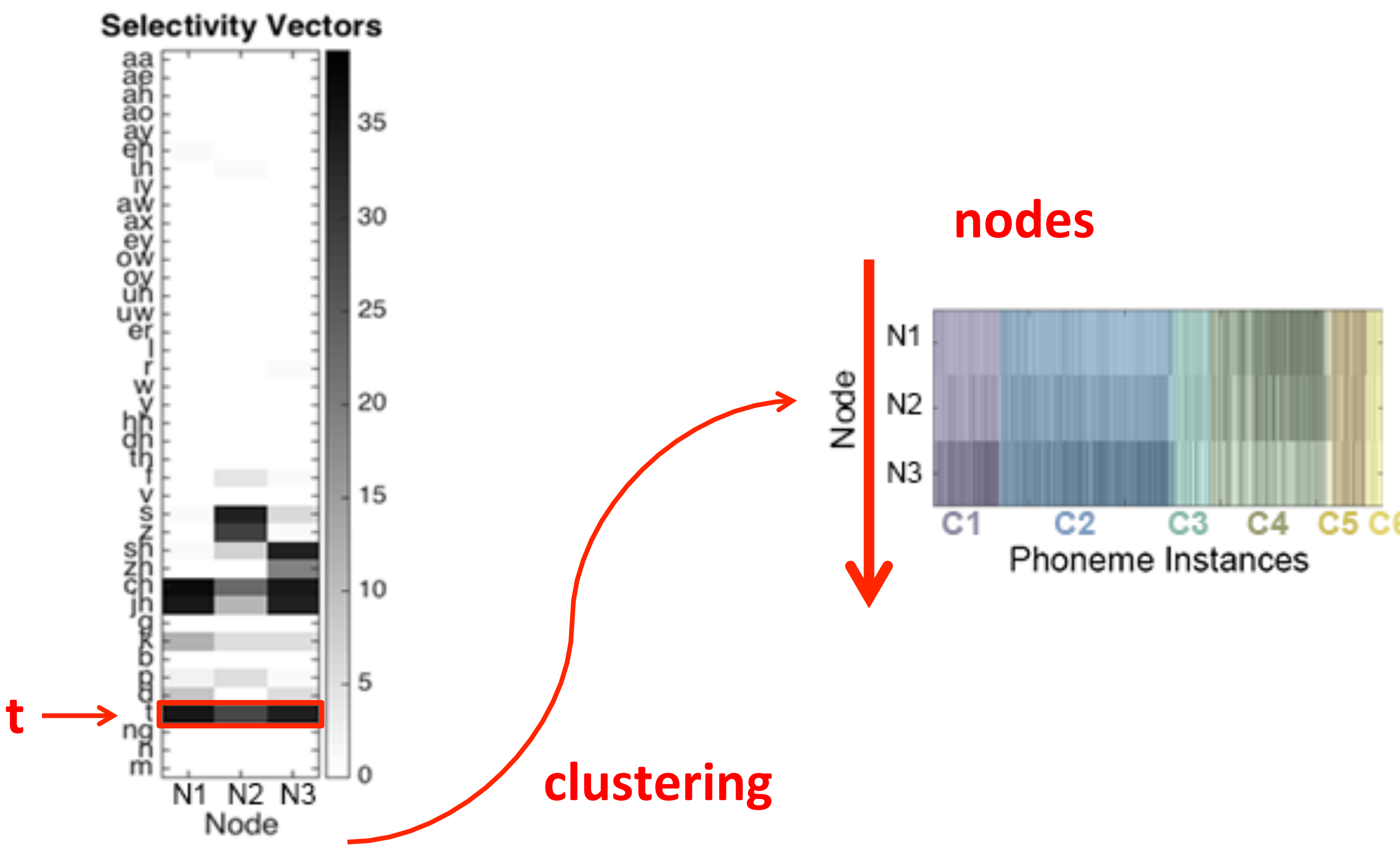
clustering

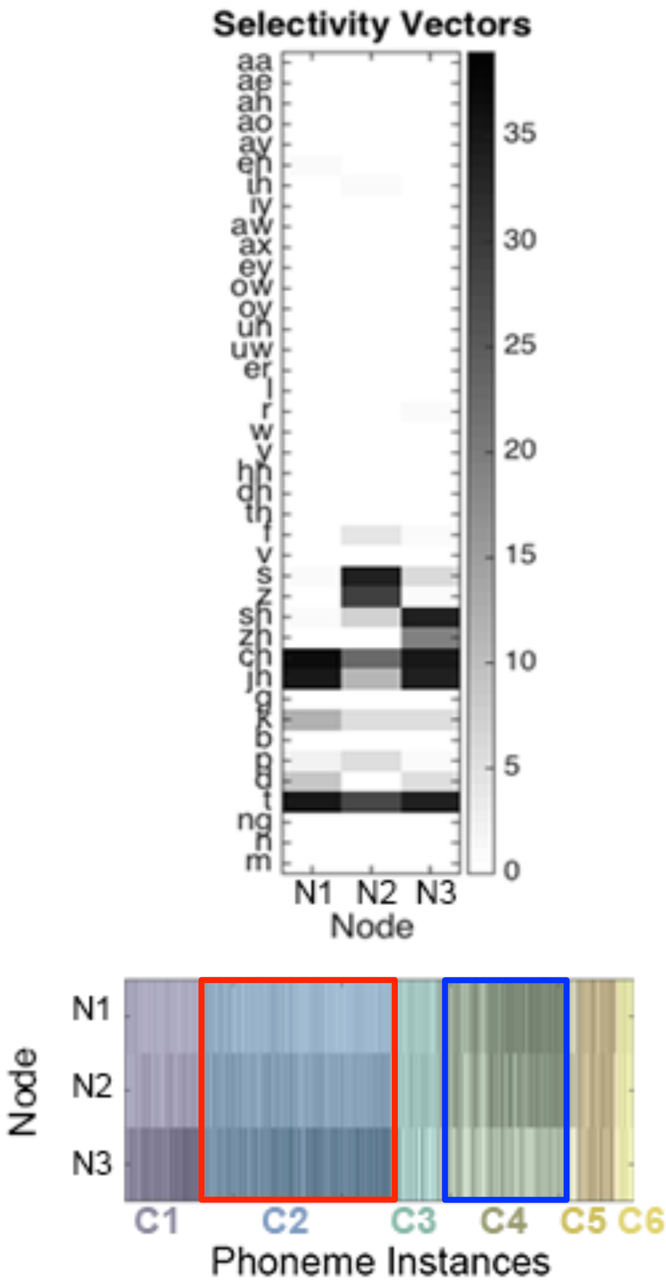
phoneme instances



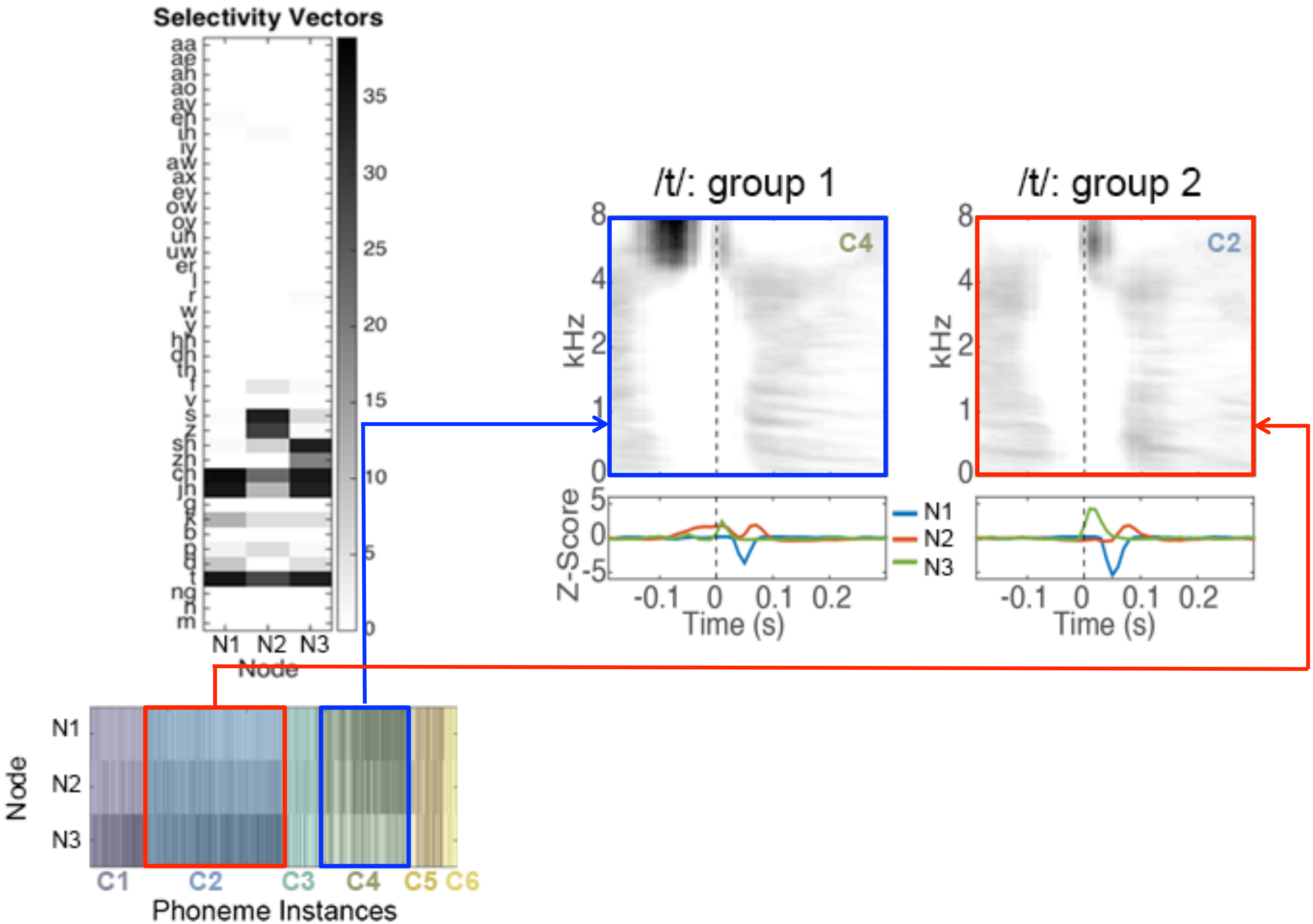
phoneme = “t”

example selectivity for three nodes (N1, N2, N3)

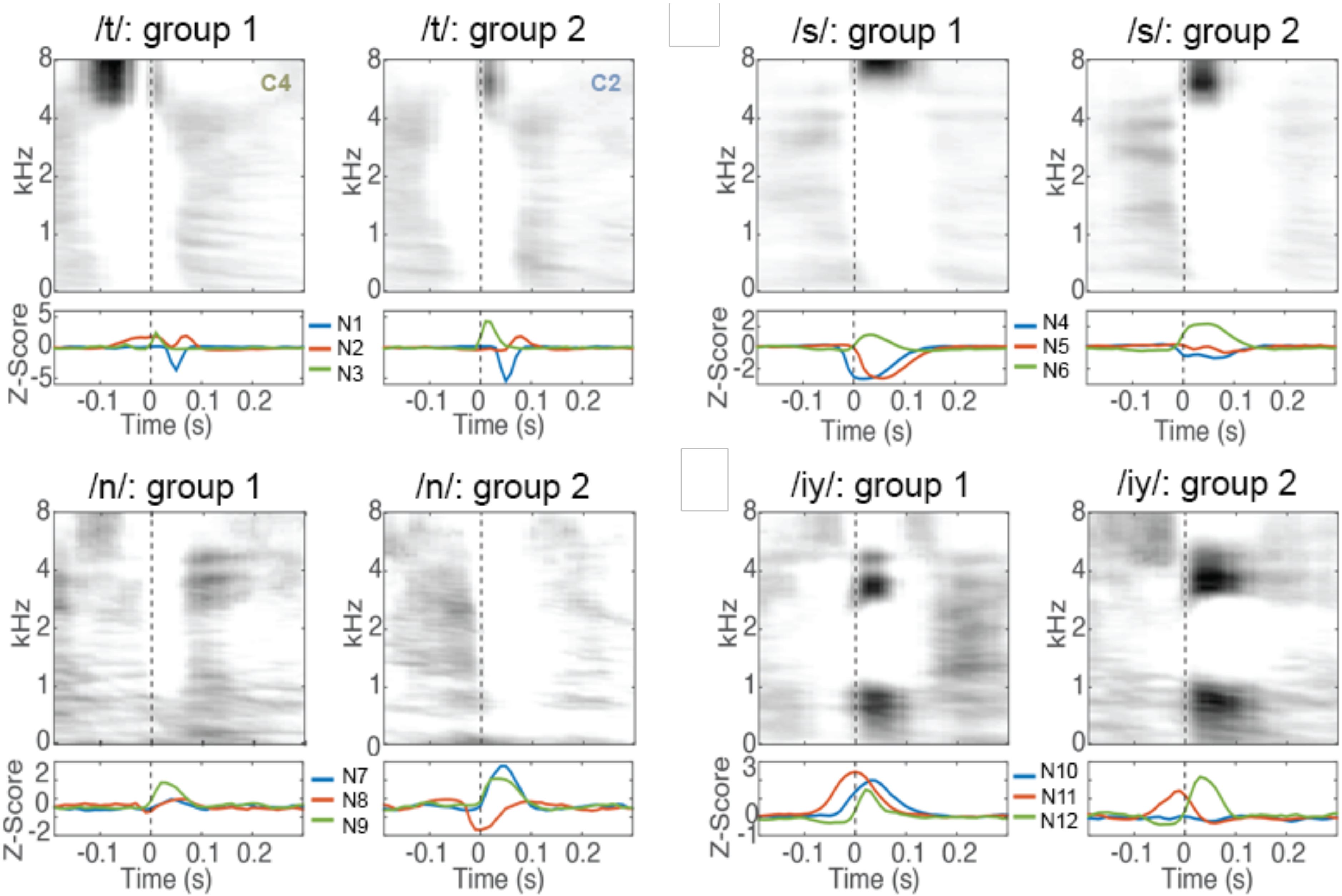




Nagamine et al. Interspeech 2015
October 14, 2015



Selectivity Vectors
aa
ae
ah
ao
ay
eh
ih
iy
aw
ax
ey
ow
oy
uh
uw
er
l
r
w
y
hh
dh
th
f
v
s
z
sh
zh
ch
jh
g
k
b
p
d
t
ng
n
m
N1 N2 N3
Node
35
30
25
20
15
10
5
0
/t/: group 1
/t/: group 2
C4
C2
kHz
8
4
2
1
0
Z-Score
5
0
-5
N1
N2
N3
-0.1 0 0.1 0.2
Time (s)
Node
N1
N2
N3
C1 C2 C3 C4 C5 C6
Phoneme Instances


/t/: group 1
/t/: group 2
/s/: group 1
/s/: group 2
C4
C2
kHz
Z-Score
Time (s)
N1
N2
N3
N4
N5
N6
/n/: group 1
/n/: group 2
/iy/: group 1
/iy/: group 2
N7
N8
N9
N10
N11
N12

Summary of findings

1. Single nodes and populations of nodes in a layer are selective to phonetic features

2. Node selectivity to phonetic features becomes more explicit in deeper layers

3. Network invariance is learned through explicit representation of sources of variability

Questions?