Emergent Gestural Scores in a Recurrent Neural Network Model of Vowel Harmony

Society for Computation in Linguistics
February 17, 2021

Caitlin Smith[1], Charlie O'Hara[2], Eric Rosen[1], Paul Smolensky[1,3]

[1]Johns Hopkins University
[2]University of Southern California
[3]Microsoft Research

JOHNS HOPKINS UNIVERSITY

USC

Modeling Phonology and Phonetics with a Recurrent Neural Network

Underlying Form → **Phonology** → Surface Form → **Phonetics** → Articulation

Recurrent neural networks compute phonological surface forms from underlying forms (Hare 1990; Prickett 2019)

Recurrent neural networks compute articulatory trajectories from strings of segments (Jordan 1986; Biasutto-Lervat & Ouni 2018)

Modeling the Phonology-Phonetics Interface with a Recurrent Neural Network

- Can a recurrent neural network learn to compute articulatory trajectories directly from input phonological segments without being provided any intermediate linguistic structure?

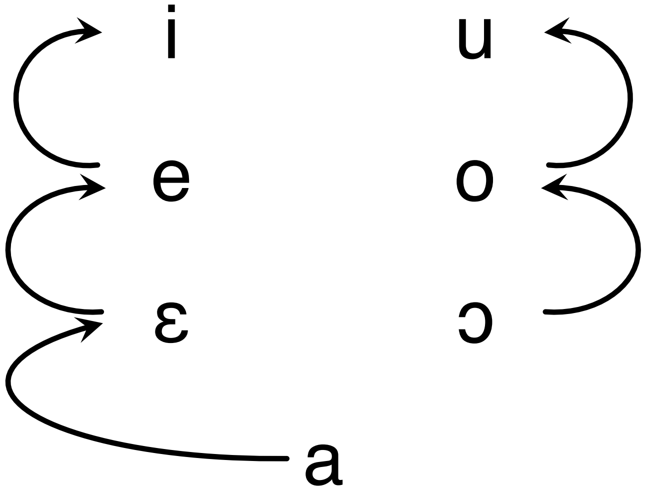
- If so, when tasked with learning a pattern of phonological alternation (e.g. vowel harmony), how does the network represent and generate the pattern?

GestNet: encoder-decoder network that generates articulatory trajectories from string of phonological input segments

Nzebi Stepwise Height Harmony

(Guthrie 1968, Clements 1991, Parkinson 1996, Kirchner 1996, Smith 2020)

In presence of trigger /-i/, each nonhigh vowel raises one ‘step’ along a height scale

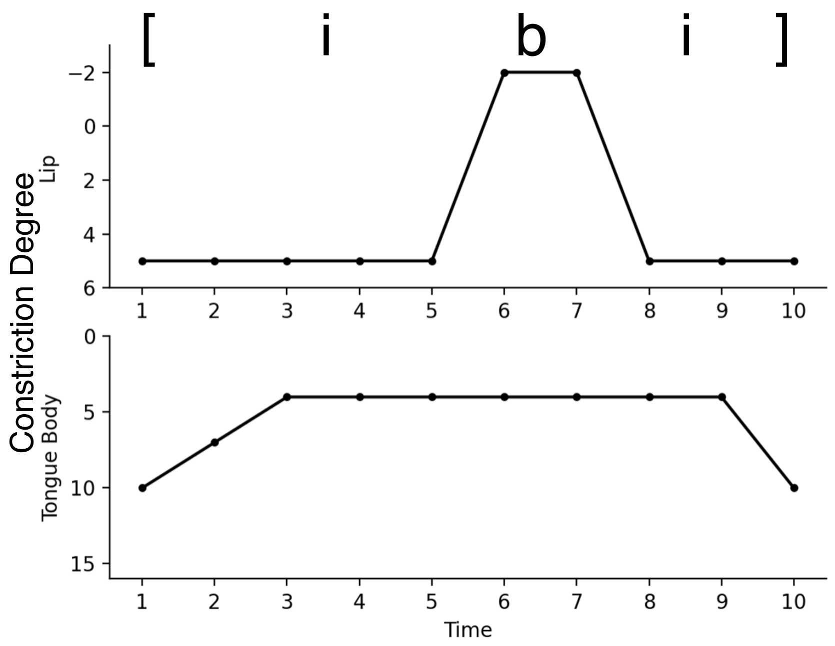


Non-Raising Context	Raising Context	Gloss
[betə]	[bit-i]	‘carry’
[βoːmə]	[βuːm-i]	'breathe’
[sɛbə]	[seb-i]	‘laugh’
[mɔnə]	[mon-i]	‘see’
[salə]	[sɛl-i]	‘work’

Modeling the Phonology-Phonetics Interface with a Recurrent Neural Network

Segments → **GestNet** encoder-decoder recurrent neural network → Articulatory Trajectories

/e b i/

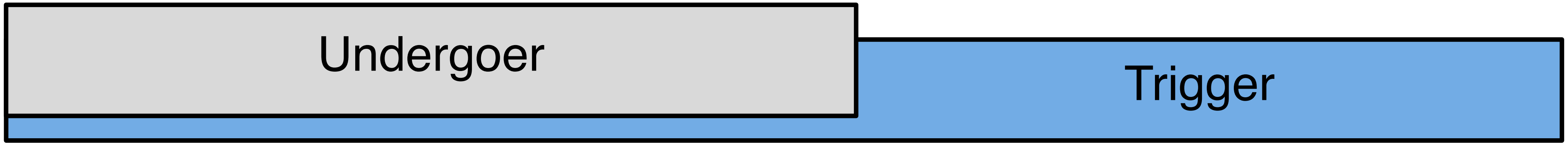


Proposal:
GestNet develops emergent structure analogous to the abstract representations of the Gestural Harmony Model

Representing Harmony with Gestures

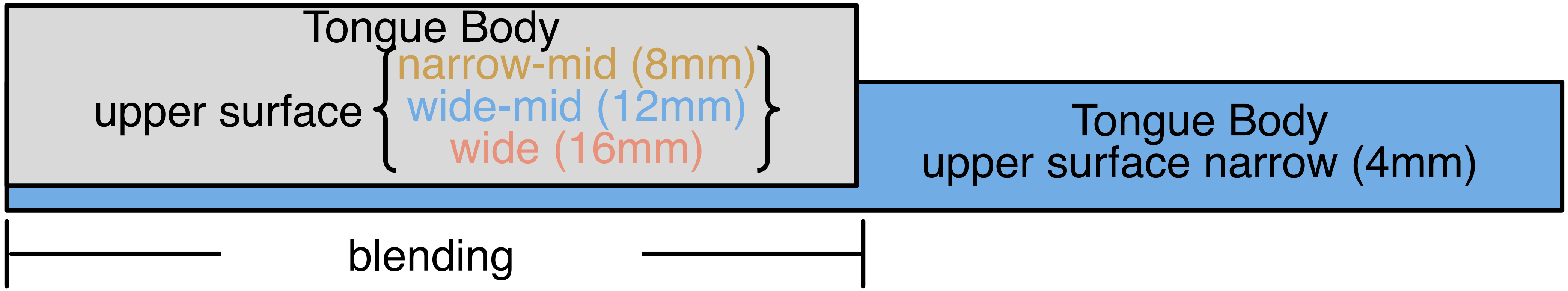
- Articulatory Phonology (Browman & Goldstein 1986, 1989):
  - Dynamically-defined, goal-based units of phonological representation
  - Specified for target articulatory state (e.g. labial closure)

- Gestural Harmony Model (Smith 2016, 2018): harmony-triggering gesture extends to overlap gestures of other segments in a word (undergoers)

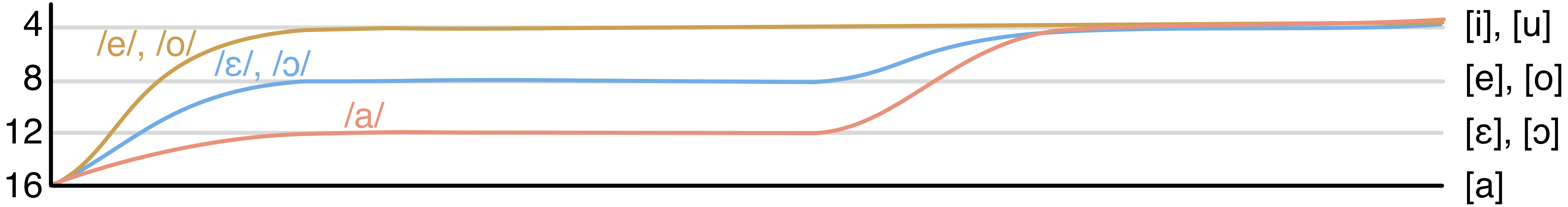


A Gestural Analysis of Nzebi
(Smith 2020)

Vowel raising harmony due to overlap by upper surface narrowing gesture of suffix high vowel /i/



Resulting tongue body/upper surface aperture (mm):



Modeling the Phonology-Phonetics Interface with a Recurrent Neural Network

Segments

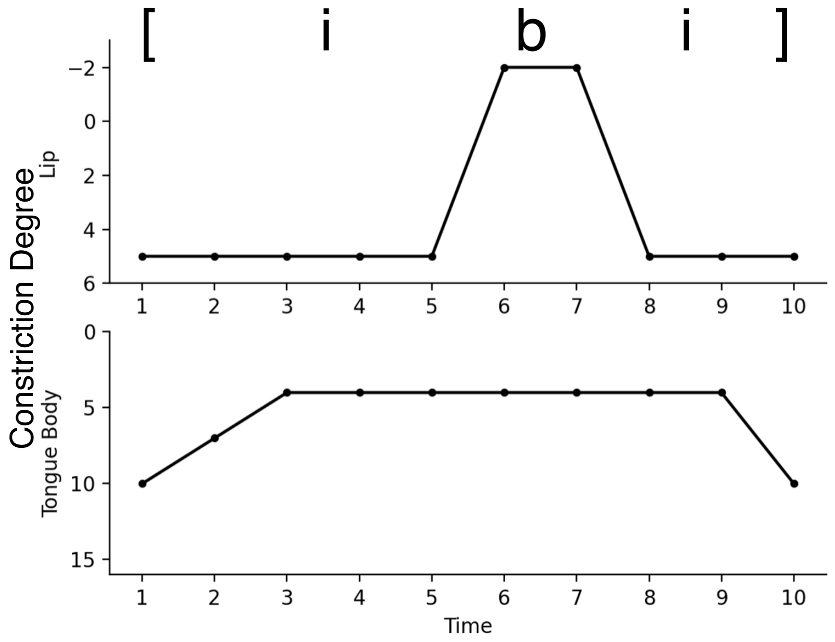
/e b i/

→

GestNet
encoder-decoder
recurrent neural network

→

Articulatory Trajectories

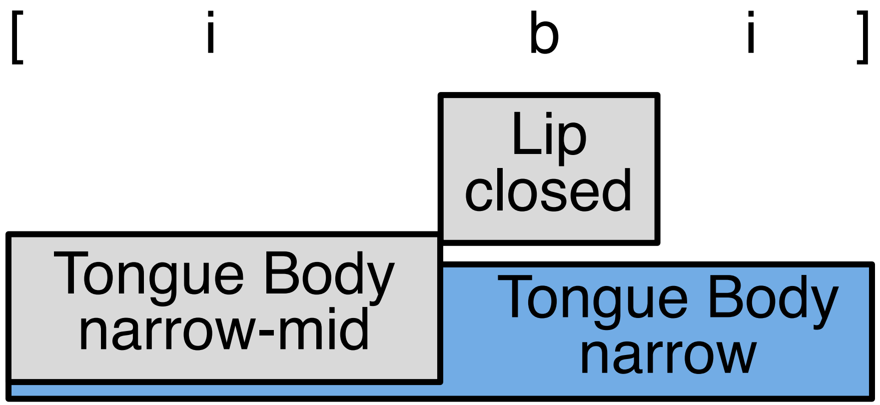


Modeling the Phonology-Phonetics Interface in Gestural Phonology

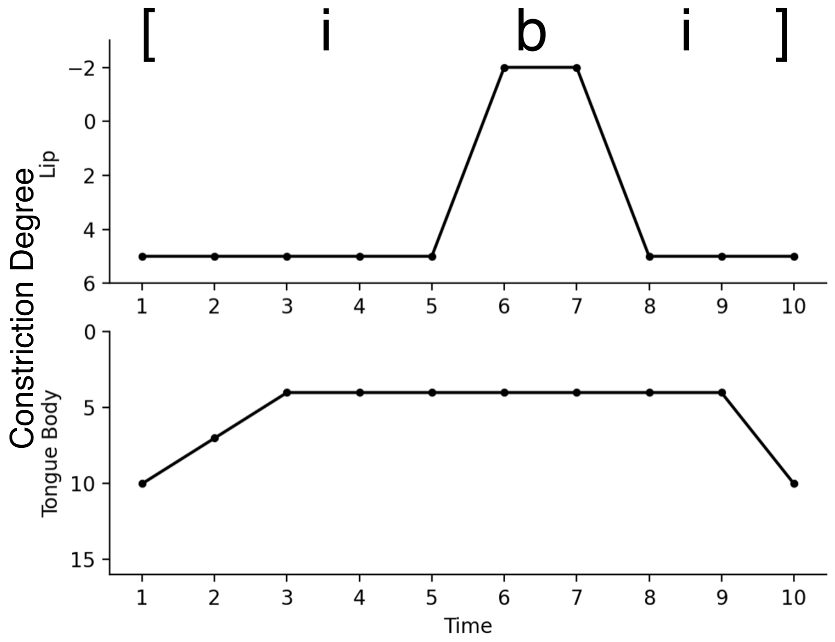
Segments

/e b i/

Gestural Score



Articulatory Trajectories

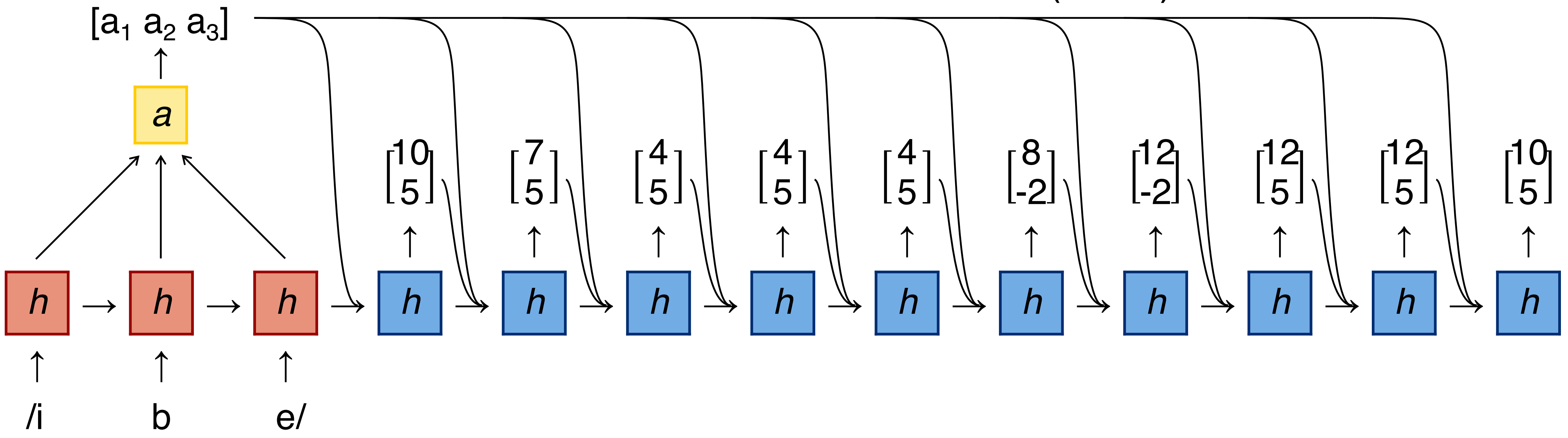


GestNet

GestNet's Encoder-Decoder Architecture

(Cho et al. 2014; Sutskever et al. 2014; Bahdanau et al. 2015; Luong et al. 2015)

Attention (a): provide each decoder hidden state (blue h) with access to all encoder hidden states (red h)



Encoder: process one input vector at each time step

Decoder: produce one output vector at each time step

Training the Model

Segment	Constriction Degree Target
i, u	Tongue Body 4
e, o	Tongue Body 8
ɛ, ɔ	Tongue Body 12
a	Tongue Body 16
b	Lip -2
g	Tongue Body -2

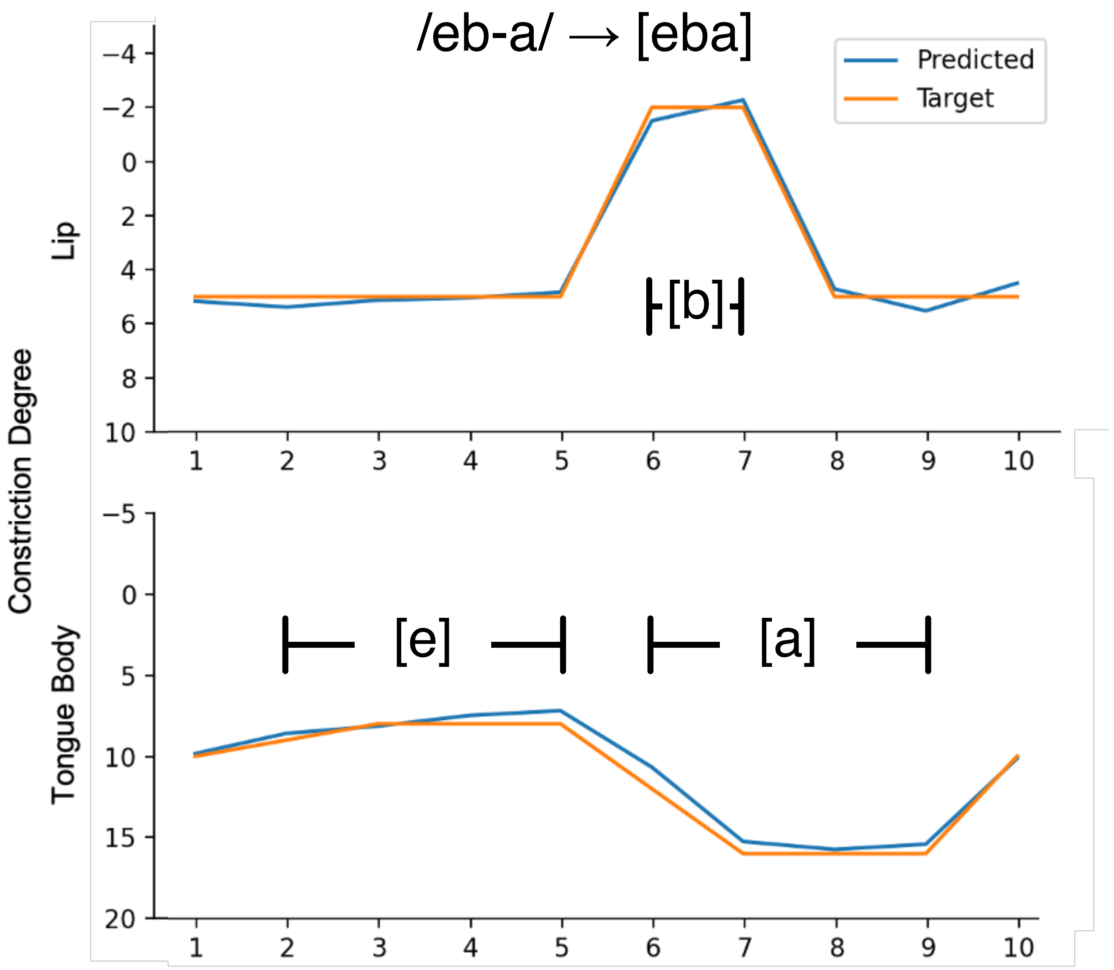
- Training data: 112 total (V)CV sequences
  - Inputs: symbols strings with C = {b, g} and V = {i, e, ɛ, a, ɔ, o, u}
  - Outputs: artificially generated trajectories for lip and tongue body positions across ten timepoints

- Height harmony pattern: In VCV in which V_2 is high vowel /i/ or /u/, V_1 undergoes one-step raising (i.e. /eb-a/→[eba] but /eb-i/→[ibi])

- Trained twenty models for 200 epochs each

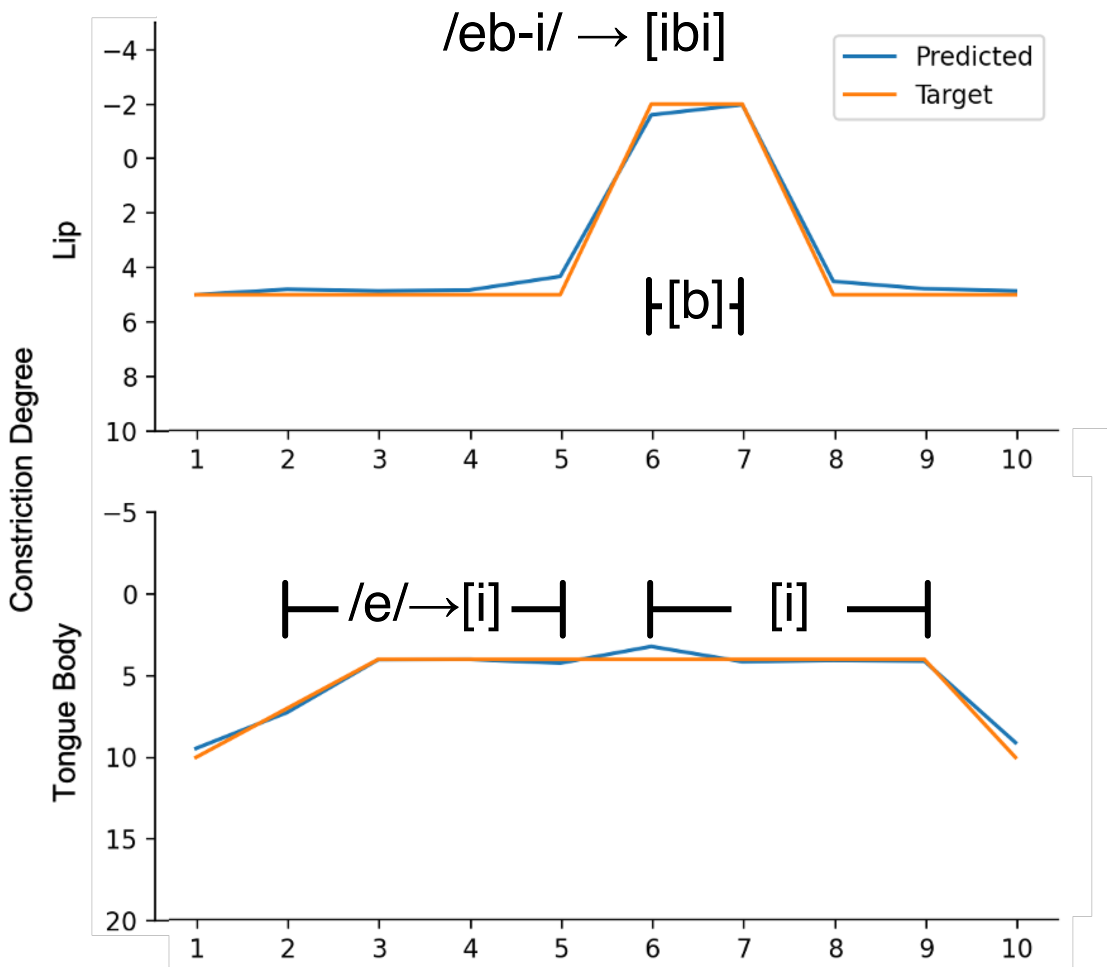
Results & Analysis

Model Accuracy



- All models produced highly accurate lip and tongue body trajectories for VCV sequences after training
- V_1 produced without raising before non-high vowels
- V_1 produced with one-step raising before high vowels

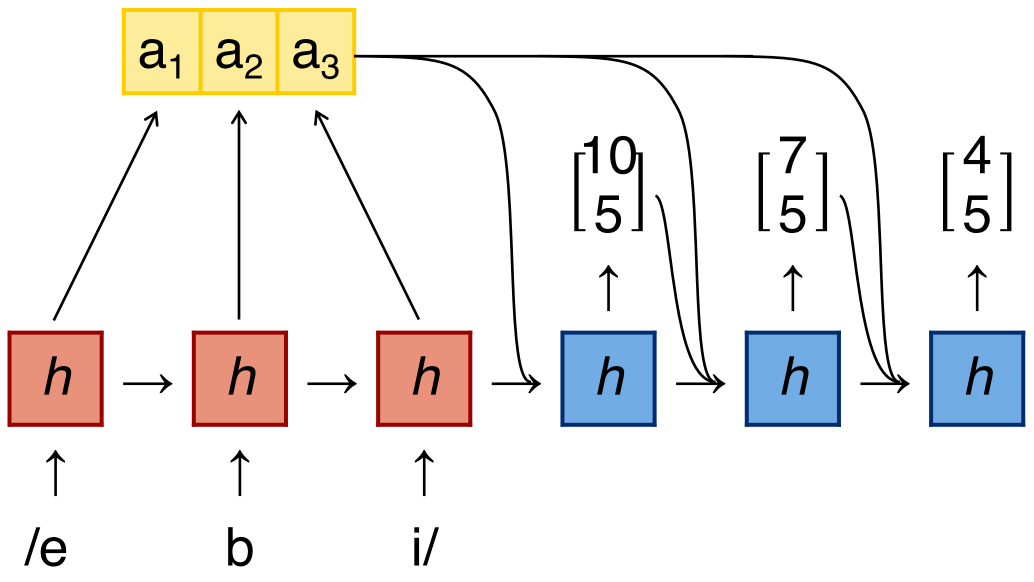
Model Accuracy



- All models produced highly accurate lip and tongue body trajectories for VCV sequences after training
- V_1 produced without raising before non-high vowels
- V_1 produced with one-step raising before high vowels

What are our models learning when they learn to produce these patterns?

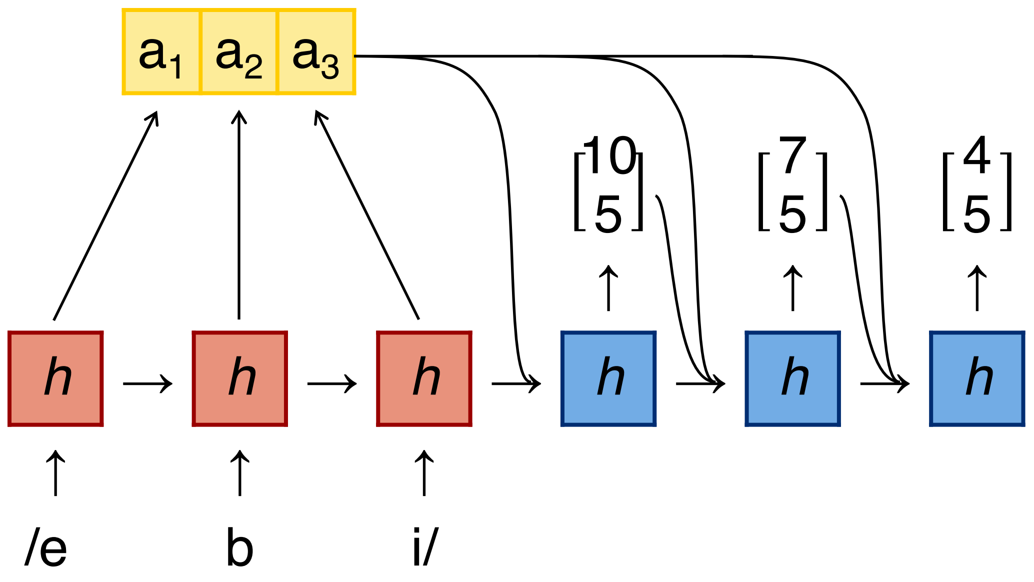
Examining Encoder-Decoder Attention



- Encoder-decoder attention provides simple recurrent neural networks with short memories a way to look back to encoder hidden states

- Degree of attention paid to an encoder hidden state can be used as measure of how much influence an input segment has on output at specific timepoint

Examining Encoder-Decoder Attention

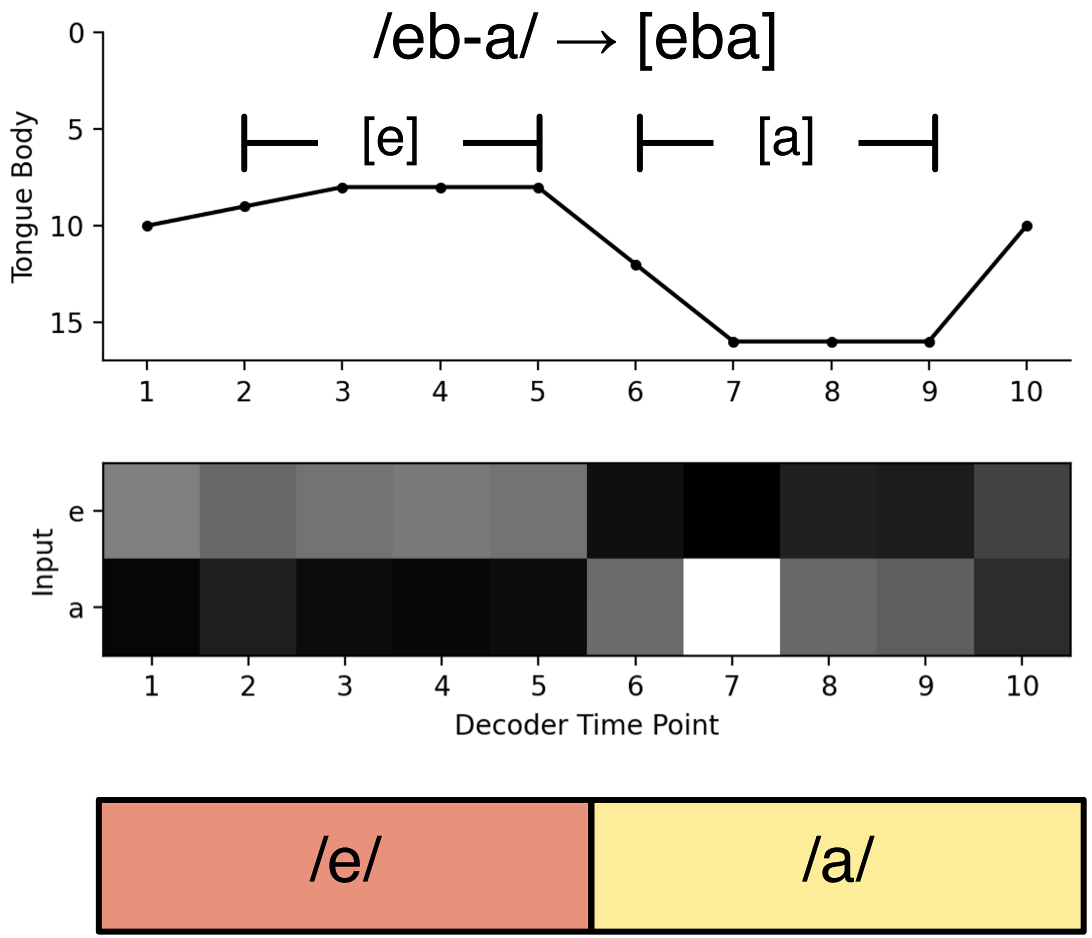


Proposal: Patterns of encoder-decoder attention reflect patterns of gestural activation in a word's gestural score

- Effective attention: attention weight multiplied by magnitude of its encoder hidden state vector

- At each decoder timepoint, record vector of effective attention weights to determine degree to which how much or how little each encoder hidden state affects the decoder hidden state

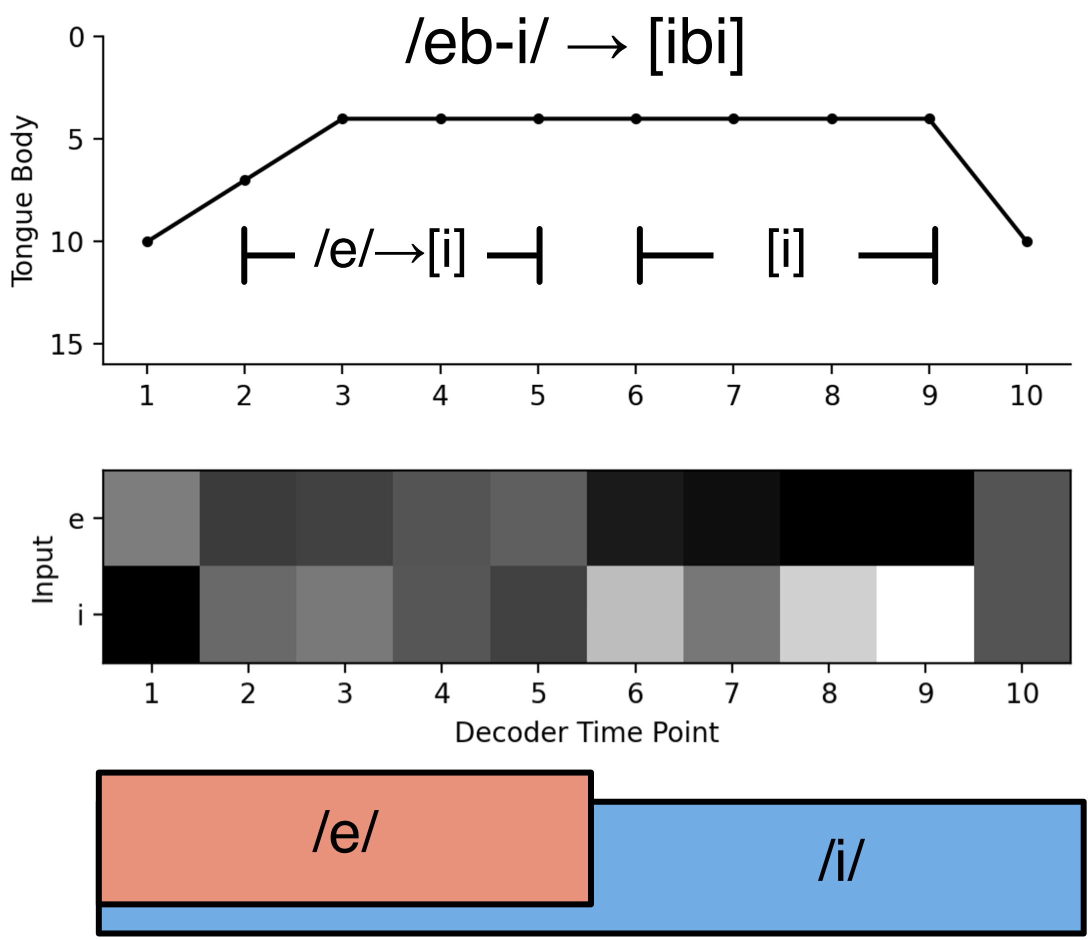
Attention Maps: Qualitative Analysis



Lighter color = more attention

- Attention maps show how much the model's decoder attends to each input segment at each time point
- Non-triggering V_2: V_1 and V_2 each receive attention during their own productions, but not while the other is being produced
- Consistent with sequential gestural activation

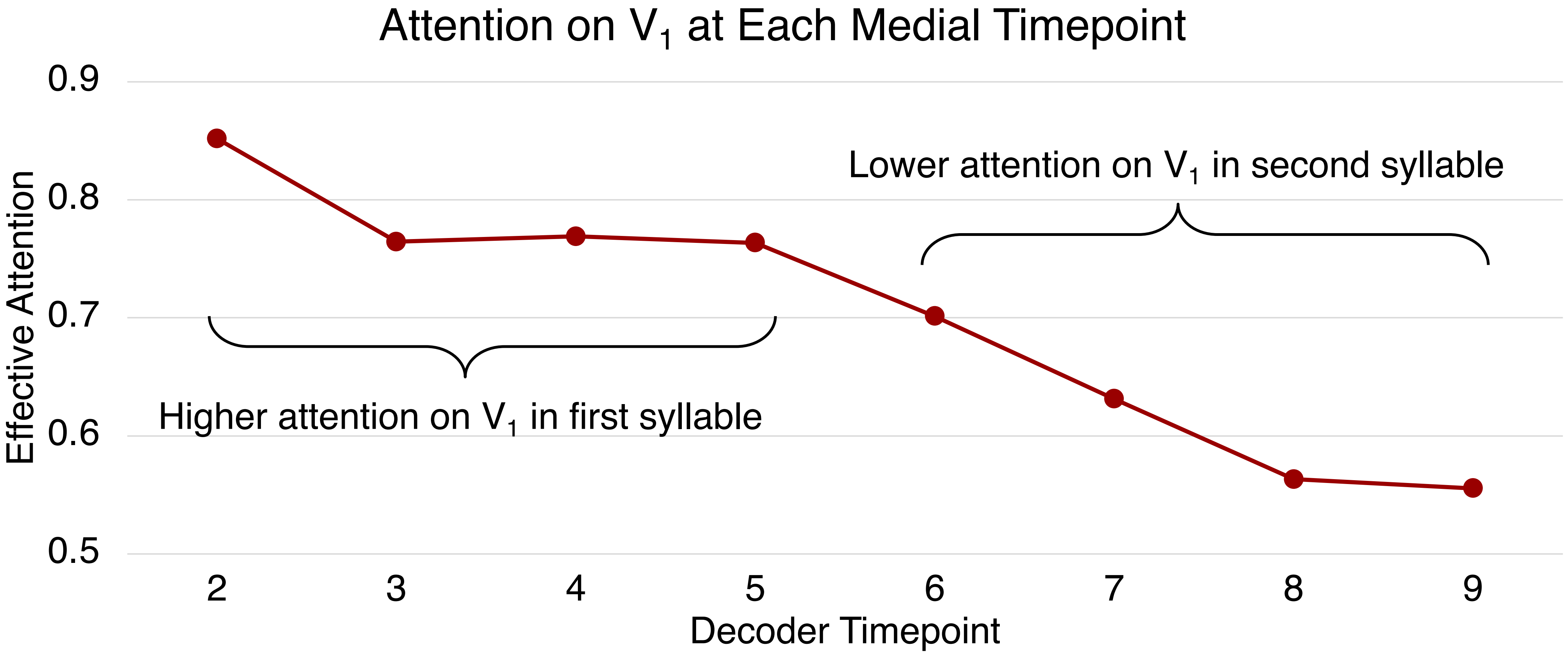
Attention Maps: Qualitative Analysis



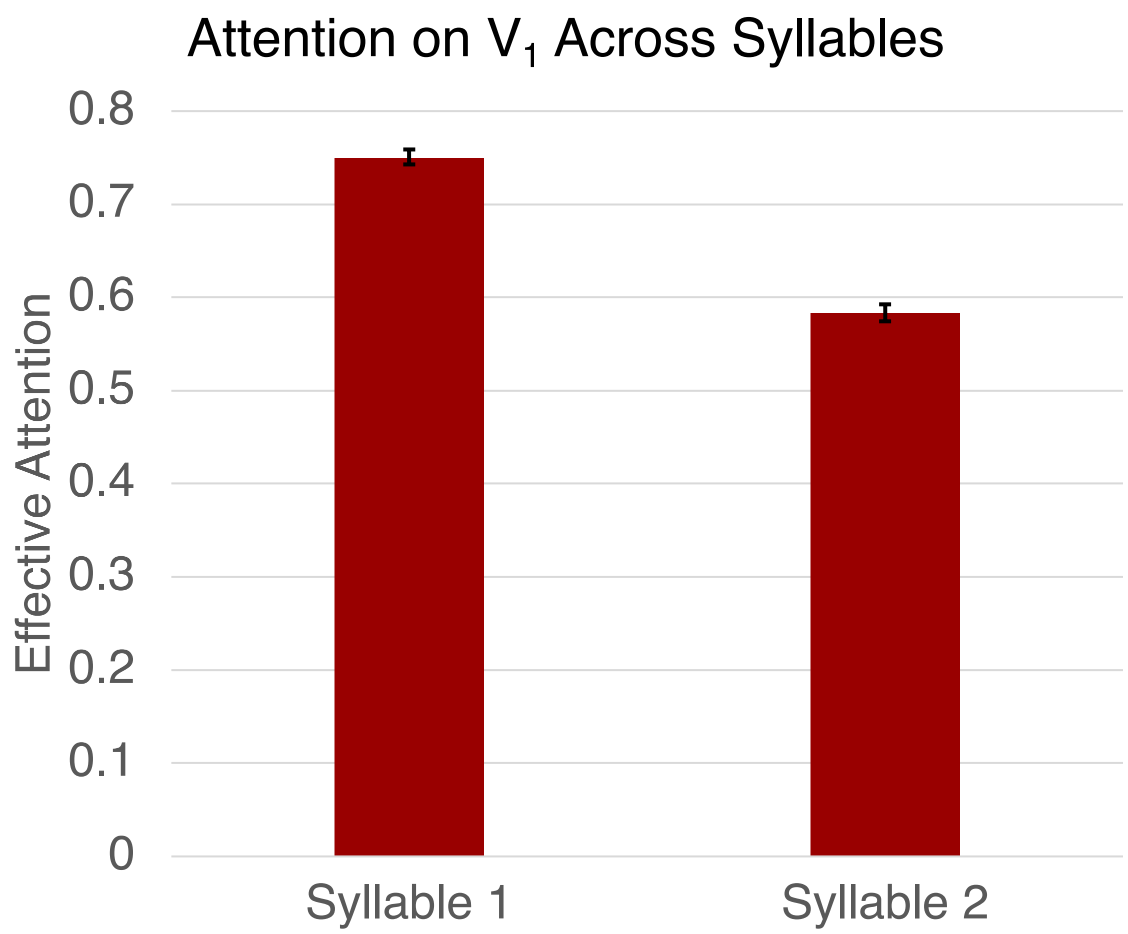
Lighter color = more attention

- Attention maps show how much the model's decoder attends to each input segment at each time point
- Triggering V_2:
  - V_1 receives attention during first half of word
  - V_2 receives attention throughout the entire word
- Consistent with overlapping gestural activation

Attention Maps: Quantitative Analysis

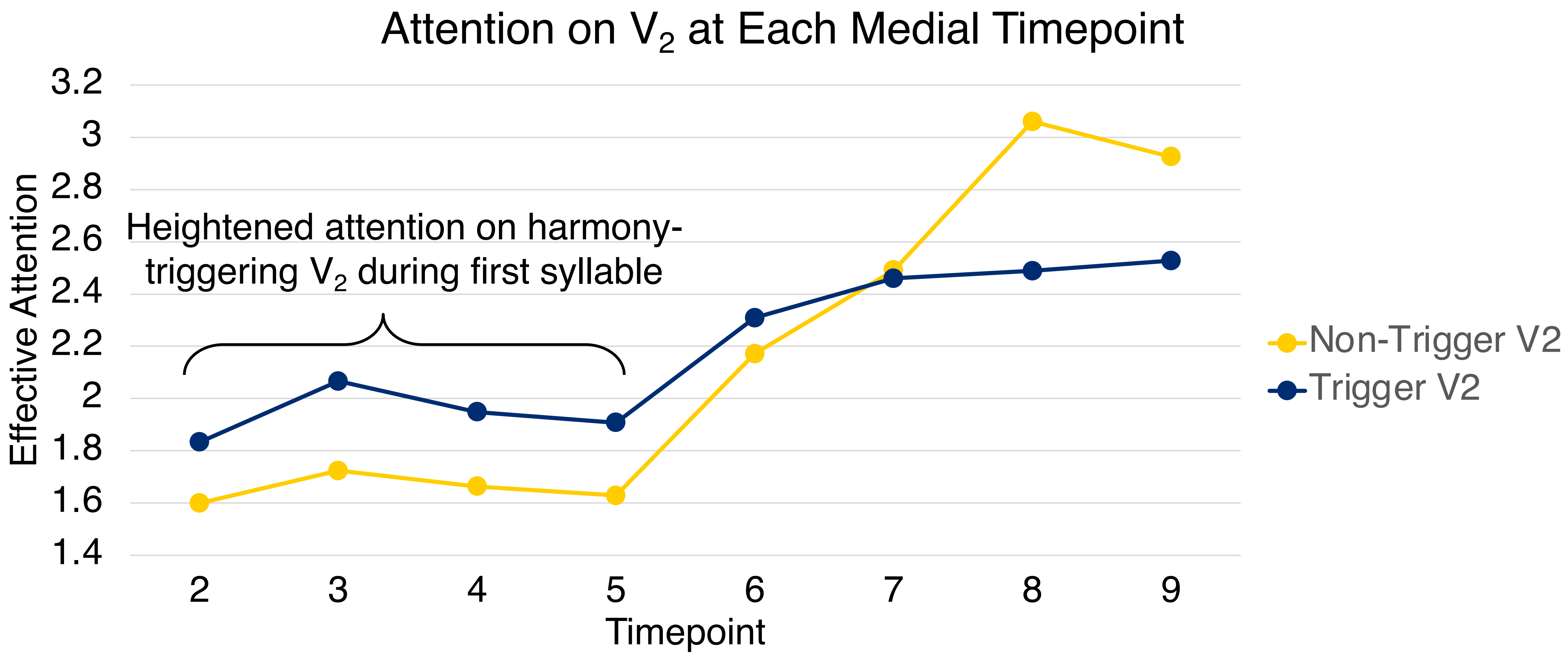


Attention Maps: Quantitative Analysis

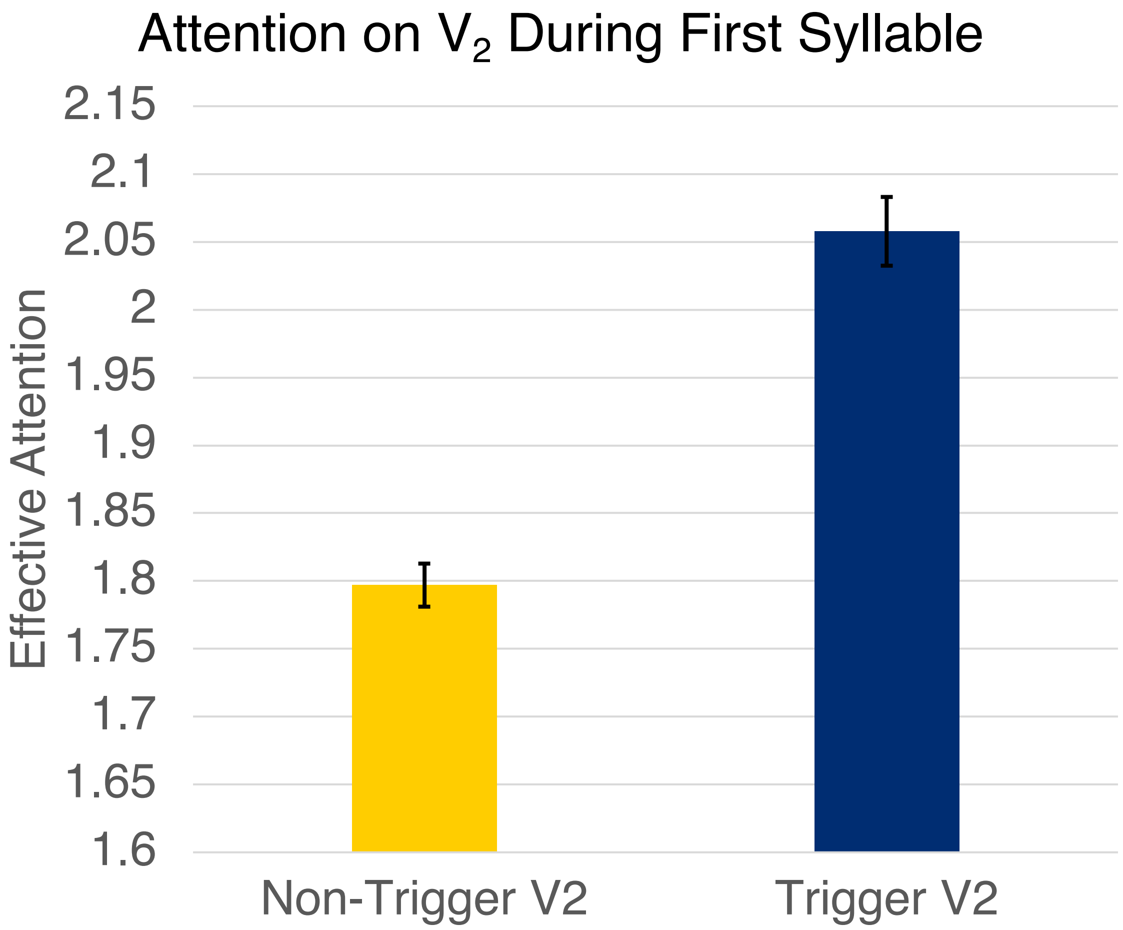


- Mixed effects model confirms these attention patterns are significant
- During production of first syllable (decoder timepoints 2-5), V_1 input segment receives significantly more attention than during production of second syllable (decoder timepoints 6-9) ($p < 0.001$)
- Gesture of V_1 is active during first syllable and not active during second syllable

Attention Maps: Quantitative Analysis



Attention Maps: Quantitative Analysis



- Mixed effects model confirms these attention patterns are significant
- During production of first syllable (timepoints 2-5), harmony-triggering V_2 input segment receives significantly more attention than non-triggering V_2 ($p < 0.001$)
- Gesture of harmony-triggering V_2 is active during first syllable; gesture of non-triggering V_2 is not

Conclusion

Conclusion

- GestNet models reliably learn a pattern of stepwise height harmony
- Models develop emergent structure analogous to the abstract representations of gestural phonology
- Patterns of encoder-decoder attention are consistent with patterns of gestural activation assumed in the Gestural Harmony Model
- Next steps: additional model analysis, additional phonological patterns