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

Caitlin Smith¹, Charlie O’Hara², Eric Rosen¹, Paul Smolensky¹,³

¹Johns Hopkins University
²University of Southern California
³Microsoft Research
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).
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’   |
| [βɔ:meə]            | [βuːm-i]        | ‘breathe’ |
| [sɛbəə]             | [seb-i]         | ‘laugh’   |
| [mɔnəə]             | [mon-i]         | ‘see’     |
| [sælə]              | [sɛl-i]         | ‘work’    |
Modeling the Phonology-Phonetics Interface with a Recurrent Neural Network

Segments
/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)

[Diagram of Trigger and Undergoer]
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):

[4, 8, 12, 16]

[i], [u]
[e], [o]
[ɛ], [ɔ]
[a]
Modeling the Phonology-Phonetics Interface with a Recurrent Neural Network

Segments
/e b i/

GestNet
encoder-decoder
recurrent neural network

Articulatory Trajectories

Constriction Degree
Modeling the Phonology-Phonetics Interface in Gestural Phonology

Segments /e b i/

Gestural Score

Tongue Body narrow-mid

Tongue Body narrow

Lip closed

Articulatory Trajectories

Constriction Degree

Lip
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

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

| Segment | Constriction Degree Target |
|---------|----------------------------|
| i, u    | Tongue Body 4              |
| e, o    | Tongue Body 8              |
| \varepsilon, \sigma | Tongue Body 12 |
| a       | Tongue Body 16              |
| b       | Lip -2                     |
| g       | Tongue Body -2             |

- 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
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.
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

- 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

Proposal: Patterns of encoder-decoder attention reflect patterns of gestural activation in a word’s gestural score
Attention Maps: Qualitative Analysis

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.

Input

Decoder Time Point

Lighter color = more attention

/e/ /a/

/eb-a/ → [eba]
Attention Maps: Qualitative Analysis

- 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 on $V_1$ at Each Medial Timepoint

- Higher attention on $V_1$ in first syllable
- Lower attention on $V_1$ in second syllable
Attention Maps: Quantitative Analysis

- Mixed effects model confirms these attention patterns are significant

- During production of first syllable (decoder timepoints 2-5), V₁ input segment receives significantly more attention than during production of second syllable (decoder timepoints 6-9) \( (p < 0.001) \)

- Gesture of V₁ is active during first syllable and not active during second syllable
Attention Maps: Quantitative Analysis

Attention on V₂ at Each Medial Timepoint

- Heightened attention on harmony-triggering V₂ during first syllable

**Effective Attention**

- Non-Trigger V₂
- Trigger V₂
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