Proceedings of the Society for Computation in Linguistics

Volume 4 Article 7

2021

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

Caitlin Smith
Johns Hopkins University, csmit372@jhu.edu

Charlie O'Hara
University of Southern California, charleso@usc.edu

Eric Rosen
Johns Hopkins University, erosen27@jhu.edu

Paul Smolensky
Johns Hopkins University, psmolen1@jhu.edu

Recommended Citation

Smith, Caitlin; O'Hara, Charlie; Rosen, Eric; and Smolensky, Paul (2021) "Emergent Gestural Scores in a Recurrent Neural Network Model of Vowel Harmony," *Proceedings of the Society for Computation in Linguistics*: Vol. 4 , Article 7.
DOI: https://doi.org/10.7275/qyey-4j04
Available at: https://scholarworks.umass.edu/scil/vol4/iss1/7

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

Caitlin Smith
Department of Cognitive Science
Johns Hopkins University
csmit372@jhu.edu

Charlie O'Hara
Department of Linguistics
University of Southern California
charleso@usc.edu

Eric Rosen
Department of Cognitive Science
Johns Hopkins University
erosen27@jhu.edu

Paul Smolensky
Department of Cognitive Science
Johns Hopkins University
Microsoft Research AI
psmolen1@jhu.edu

Abstract

In this paper, we present the results of neural network modeling of speech production. We introduce GestNet, a sequence-to-sequence, encoder-decoder neural network architecture in which a string of input symbols is translated into sequences of vocal tract articulator movements. We train our models to produce movements of lip and tongue body articulators consistent with a pattern of stepwise vowel height harmony. Though we provide our models with no linguistic structure, they reliably learn this harmony pattern. In addition, by probing these models we find evidence of emergent linguistic structure. Specifically, we examine patterns of encoder-decoder attention (degree of influence of specific input segments on model outputs) and find that they resemble the patterns of gestural activation assumed within the Gestural Harmony Model, a model of harmony built upon the representations of Articulatory Phonology. This result is significant as it lends support to one of the central claims of the Gestural Harmony Model: that harmony is the result of the harmony-triggering gestures extending to overlap the gestures of surrounding segments.

1 Introduction

In partial height harmony, some undergoer vowels may assimilate only partially to a trigger vowel, approaching the trigger's height without matching it. Partial height harmonies often proceed in a stepwise, or chain-shifting, fashion, with each vowel raising one step along a height scale. This is illustrated by the vowel raising harmony of Nzebi, a Bantu language of Gabon (Guthrie, 1968; Kirchner, 1996; Parkinson, 1996; Smith, 2020b). In Nzebi, the suffix /-i/ occurs immediately after verb roots in some tenses and triggers one-step raising of preceding root vowels. Before this harmony triggering suffix, high-mid vowels /e/ and /o/ surface as [i] and [u], respectively; low-mid vowels /ɛ/ and /ɔ/ surface as [e] and [o], respectively; and low /a/ surfaces as [ɛ]. This is illustrated by the data in (1).

(1) Root vowels in non-raising vs. raising contexts

a.	[bɛ̰t]	[bḭt-i]	'carry'
b.	[βo̰ːm]	[βṵːm-i]	'breathe'
c.	[sɛ̰b]	[sḛb-i]	'laugh'
d.	[mɔ̰n]	[mo̰n-i]	'see'
e.	[sa̰l]	[sɛ̰l-i]	'work'

This pattern of stepwise vowel raising is summarized in Figure 1.

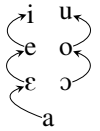

Figure 1: Pattern of vowel raising in the stepwise height harmony of Nzebi

Smith (2020b) provides an analysis of Nzebi's stepwise height harmony within the Gestural Harmony Model (Smith, 2016, 2018), a theory of vowel and vowel-consonant harmony couched within the framework of Articulatory Phonology (Browman and Goldstein, 1986, 1989). In this model, harmony is analyzed as the result of the extension of a harmony-triggering gesture such that it overlaps the gestures of surrounding segments.

In order to test the validity of this analysis, in this

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 61-70.
Held on-line February 14-19, 2021

paper we introduce GestNet, a type of sequence-to-sequence, encoder-decoder neural network architecture, and use it to model this stepwise height harmony pattern.[1] GestNet takes as its inputs sequences of underlying phonological symbols and outputs sequences of vocal tract articulator movements. By providing our models with no linguistic structure and probing their internal states, we find emergent structure consistent with the linguistic analysis of height harmony within the Gestural Harmony Model.

The current work also contributes to the growing body of research that uses recurrent neural networks to model aspects of phonology and its interfaces. A large portion of this research has primarily involved using recurrent neural networks as phonotactic models by performing language modeling tasks over strings of segments (Elman, 1990; Rodd, 1997; Silfverberg et al., 2018; Mirea and Bicknell, 2019; Mayer and Nelson, 2020; Rosen, 2021). Comparatively fewer models have been presented that perform sequence-to-sequence mapping between underlying and surface phonological forms. Gaskell et al. (1995) use a simple recurrent network to model mappings from surface forms to underlying forms of words exhibiting consonant place assimilation. Prickett (2019) uses an encoder-decoder model to map from underlying to surface phonological forms that have undergone various derivationally transparent and opaque processes. Models that incorporate gestural, rather than featural, phonological forms are even less common. A recent example of such work comes from Tilsen (2020), who presents a model that maps articulatory trajectories collected using electromagnetic articulography to patterns of gestural activation. By contrast, GestNet is designed to map directly from an input string of phonemes to an output string of articulatory trajectories.

The paper is organized as follows. Section 2 introduces the Gestural Harmony Model and summarizes the analysis of Nzebi height harmony within that model. Section 3 outlines our methods for constructing neural network models of height harmony, including the data we used to train our models, the architecture of our GestNet models, and our training procedures. Section 4 presents the results of training, including model performance and interpretability. Section 5 concludes.

[1]The code for GestNet, as well as our training data, can be found at https://github.com/caitlinsmith14/gestnet.

2 Height Harmony in the Gestural Harmony Model

2.1 The Gestural Harmony Model

Gestures are the units of sub-segmental representation assumed within the framework of Articulatory Phonology (Browman and Goldstein, 1986, 1989). They are dynamically-defined, goal-based units, with each gesture being specified for a target articulatory state to be achieved during its period of activation. This target state is specified in terms of a primary articulator, a constriction location, and a constriction degree. The constriction location of a gesture is specified as a point or region along the static surface of the vocal tract, while constriction degree refers to the aperture of the constriction between the primary articulator and the constriction location. For instance, a gesture for the high vowel /i/ can be specified as having a narrow constriction between the tongue body and the upper surface of the vocal tract as its target articulatory state. A number of additional parameters determine precisely how and when a gesture achieves its target articulatory state; for reasons of space, they are omitted from our discussion.

In gestural phonology, forms are often displayed in a gestural score such as the one in Figure 2 for a VCV sequence. In a gestural score, the activation periods of vowel gestures are typically sequential, with the second vowel in a sequence activating once the previous vowel has deactivated. A vowel gesture and the gesture of its onset consonant, meanwhile, typically activate synchronously.

Figure 2: Gestural score for a VCV sequence

The Gestural Harmony Model (Smith, 2016, 2018) adopts many aspects of the gestural representations of Articulatory Phonology. In this model, harmony is the result of a gesture extending its period of activation to overlap the gestures of preceding and/or following segments. A persistent, or non-self-deactivating, gesture is one that does not deactivate when its target articulatory state is reached, but rather remains active and extends to overlap the gestures of following segments. An anticipatory, or early-activating, gesture is one that activates before its scheduled starting point, extend-

ing to overlap the gestures of preceding segments. In the Gestural Harmony Model, harmony arises when a segment includes a gesture that is either persistent, anticipatory, or both; such a segment is a trigger of harmony. Surrounding segments undergo harmony as a result of their composite gestures being overlapped by a harmony-triggering gesture. This is illustrated by the gestural scores in Figure 3.

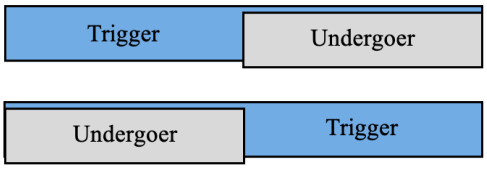

Figure 3: Harmony via overlap by a persistent gesture (above) and by an anticipatory gesture (below)

Gestural overlap often results in the concurrent activation of two gestures with antagonistic target articulatory states (e.g., narrow vs. wide constriction degree between the tongue body and the upper surface of the vocal tract). The outcome of this intergestural conflict is determined by the relative blending strengths of each of the two antagonistic gestures. According to the Task Dynamic Model of speech production (Saltzman and Munhall, 1989; Fowler and Saltzman, 1993), intergestural conflict is resolved by blending the conflicting target articulatory states of two gestures to create an intermediate target state that holds during the period of their concurrent activation. This blended target state is the weighted average of the gestures' individual target articulatory states, with the weighting in this averaging function contributed by the gestures' strength parameters, denoted α. This blending function is provided in Equation 1.

$$\frac{Target_1 \times \alpha_1 + Target_2 \times \alpha_2}{\alpha_1 + \alpha_2} \quad (1)$$

The Gestural Harmony Model appeals to the concept of blending between antagonistic gestures in order to account for cases of transparency to harmony. Smith (2020a,b) proposes an extension of the Gestural Harmony Model's analysis of transparency to cases of partial height harmony, which are analyzed as cases of partial transparency. The following section outlines such an analysis for the stepwise partial height harmony of Nzebi.

2.2 A Gestural Analysis of Nzebi

Smith (2020b) proposes an analysis of the stepwise height harmony of Nzebi within the Gestural Harmony Model. We summarize this analysis here.

Smith proposes that the partial, stepwise vowel raising harmony of Nzebi is the result of gestural blending resulting from overlap of root vowels by the anticipatory, harmony-triggering tongue body gesture of the high suffix vowel /-i/. In this analysis, the four vowel heights observed in Nzebi are represented by vowel gestures with one of four possible constriction degrees between the tongue body and the upper surface of the vocal tract: narrow (4mm), narrow-mid (8mm), wide-mid (12mm), and wide (16mm). When the narrow gesture of suffix /-i/ extends to overlap a preceding vowel with any other specified target constriction degree, it results in gestural antagonism and blending.

When overlapped by suffix /-i/, high-mid root vowels surface as high rather than resisting raising, suggesting that they have a blending strength lower than that of the triggering /-i/. This is illustrated in Figure 4, in which the first [i] in the gestural score is the result of raising.

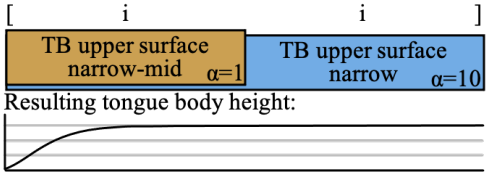

Figure 4: /e-i/ → [i-i]: blending of weak narrow-mid and strong narrow vowel gestures results in narrow tongue body aperture

Wide-mid vowels, on the other hand, raise to only an intermediate degree when overlapped by harmony-triggering /-i/, suggesting that /ɛ/ and /ɔ/ have blending strengths equal to that of /-i/. This is illustrated in Figure 5.

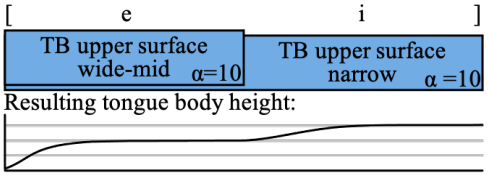

Figure 5: /ɛ-i/ → [e-i]: blending of equally strong wide-mid and narrow vowel gestures results in narrow-mid tongue body aperture

Finally, the wide vowel /a/ partially undergoes harmony and partially resists it. Because /a/ is specified for a strength that is twice the strength of the trigger gesture that overlaps it, the result

of blending is wide-mid [ɛ], closer to the intrinsic target constriction degree of wide /a/ rather than narrow /i/. This is illustrated in Figure 6.

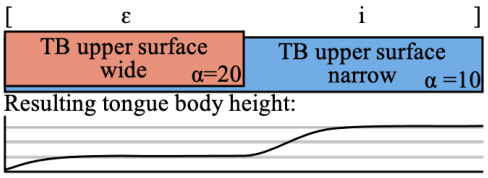

Figure 6: /a-i/ → [ɛ-i]: blending of narrow vowel gesture with stronger wide vowel gestures results in wide-mid tongue body aperture

This section has outlined the workings of the Gestural Harmony Model and the analysis of Nzebi height harmony within it. In order to test the validity of this analysis, the remainder of this paper focuses on neural network modeling of a similar pattern of stepwise height harmony.

3 Method

In this section we outline the makeup of our datasets, the architecture of our models, and the training procedures for those models.

3.1 Data

Due to the lack of a corpus of Nzebi speech, we compiled a dataset of simulated speech meant to approximate the language's stepwise height harmony pattern. The data consisted of sixteen roots with a shape of either C or VC, and seven suffixes containing a single V, as in Table 1. All Vs were taken from the set /i, e, ɛ, a, ɔ, o, u/, and consonants were taken from the set /b, g/. The full dataset comprised all possible root-suffix combinations, 112 in all.

In order to provide our models with no prior information on how many consonants and vowels should make up the target language's phonological inventory, or which phoneme was being produced in a given word, each segment of each morpheme was provided with its own unique vector embedding, which was learned throughout training. For instance, the /i/ of the roots /ib/ and /ig/ and the suffix /-i/ were all represented by separate embeddings. Likewise, the /g/ of stems /ig/, /eg/, /g/, etc. were all represented by separate embeddings.

Each root-suffix combination (CV or VCV sequence) was paired with two articulatory trajectories: one for the lip articulator and one for the tongue body articulator. Each of these articulatory

Roots		Suffixes
/ib/	/ig/	/i/
/eb/	/eg/	/e/
/ɛb/	/ɛg/	/ɛ/
/ab/	/ag/	/a/
/ɔb/	/ɔg/	/ɔ/
/ob/	/og/	/o/
/ub/	/ug/	/u/
/b/	/g/	

Table 1: Roots and suffixes in the dataset

trajectories was based on interpolation between the target constriction degrees for different segments provided by the speech synthesis toolkit TADA (Nam et al., 2004). The intrinsic target constriction degrees we assumed for each segment are provided in Table 2.[2] As a simplifying assumption, we represented articulatory trajectories as a series of ten timepoints, with the first and last timepoints each representing the neutral positions of the lip and tongue body articulators assumed before and after active speech. The medial eight timepoints were made up of the articulatory positions assumed by the two articulators during speech production. The first four active timepoints correspond to the production of the first syllable. In two-syllable words, the next four timepoints correspond to the production of the second syllable. For one-syllable words, these timepoints were padded with the values of the neutral positions of each articulator.

In order to mirror the stepwise height harmony of Nzebi, the vowels of roots preceding a high suffix vowel (either /i/ or /u/) were assigned the tongue body constriction degree associated with vowels one step higher along the height scale. For instance, in a word like /eb-i/, the output tongue body trajectory was consistent with a constriction degree of 4, consistent with narrow vowels, rather than a constriction degree of 8, during the production of the first vowel. However, the vowels of roots preceding non-high suffix vowel were assigned a tongue body constriction degree associated with the root vowel's intrinsic target.

This is illustrated in Figure 7 for the input sequence /ib-a/. The input consists of a sequence of three symbols: /i/, /b/, and /a/. The output consists of a ten-point lip aperture sequence and a ten-

[2] We abstracted away from any difference in lip constriction degree between back rounded and front unrounded vowels in order to simplify the dataset.

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

Table 2: Constriction degree targets for each segment type in the dataset

point tongue body aperture sequence. In the tongue body sequence, the sequence begins and ends with a value of 10, representing the neutral position of the tongue body. During timepoints 2 through 5, the tongue body approaches and achieves a value of 4, representing the target constriction degree for the vowel /i/. The tongue body then approaches and achieves a constriction degree value of 16, representing the target constriction degree for the vowel /a/, during timepoints 6 through 9. At timepoints 6 and 7, lip aperture is -2, corresponding to the target constriction degree of the labial consonant /b/.[3] During the rest of the lip trajectory, the lip assumes a neutral position of 5.

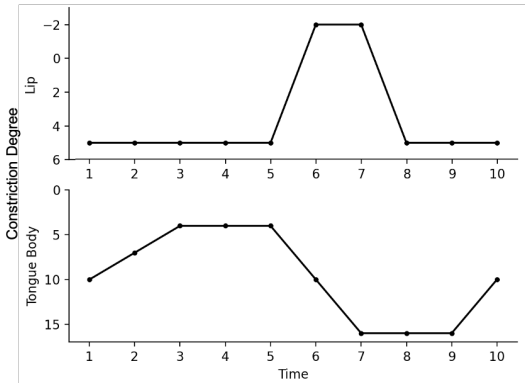

Figure 7: Sample output for sample input /ib-a/. Note that y-axes are flipped to reflect lip/tongue body height.

3.2 Model Architecture

GestNet is a type of sequence-to-sequence, or encoder-decoder, recurrent neural network (Cho et al., 2014; Sutskever et al., 2014). Encoder-decoder models were originally designed for sentence translation tasks. In our case, the task can be seen as a translation between a string of input symbols and two sequences of continuous values for articulator positions.

The role of the encoder in an encoder-decoder network is to read each input symbol one at a time, updating its hidden state at each timepoint. The final hidden state of the encoder can be thought of as containing all relevant information about the input sequence. The first hidden state of the decoder then takes in that final encoder hidden state and produces a predicted output at each time point based on the previous timepoint's hidden state and predicted output.

Due to the relatively short memory of simple recurrent neural networks, Bahdanau et al. (2015) and Luong et al. (2015) propose the mechanism of encoder-decoder attention. Rather than only passing the last encoder hidden state to the first decoder hidden state via the recurrent connection between them, encoder-decoder attention is intended to allow the decoder hidden state at any timepoint to access information contained in all encoder hidden states.

Our model architecture was implemented as follows. In the encoder, each unique segment in our dataset (37 in total), is provided with an embedding vector that is learned throughout training (Bengio et al., 2003). From there, the embedding vector is input to the hidden layer h_t, as in Equation 2.

$$h_t = tanh(W_x x_t + W_h h_{t-1} + b_h) \quad (2)$$

Both the input vector x_t and the hidden state vector of the previous time point h_{t-1} are multiplied by their respective weight matrices, W_x and W_h, summed along with the hidden state's bias terms b_h, and passed through a $tanh$ function. In the encoder, the input vector x_t is simply the embedding for the input at timepoint t. In the decoder, the definition of input vector x_t is more involved. The encoder-decoder attention mechanism is used to calculate a weighted representation of encoder hidden states, with each weight corresponding to how much attention the current decoder hidden state pays to each encoder hidden state. First, the current decoder hidden state h_t is concatenated with each encoder hidden state h_i, multiplied by its weight matrix W_a, and passed through a $tanh$ function. The resulting vector is then summed to produce the scalar a_i, as in Equation 3.

$$a_i = \sum(tanh(W_a concat(h_t, h_i) + b_a)) \quad (3)$$

[3] A negative target constriction degree is often assumed for stop consonants in Articulatory Phonology. While the achievement of such a constriction degree is of course not physically possible, this 'virtual target' allows models of speech to achieve the kinematics and tight closure consistent with the production of stop consonants.

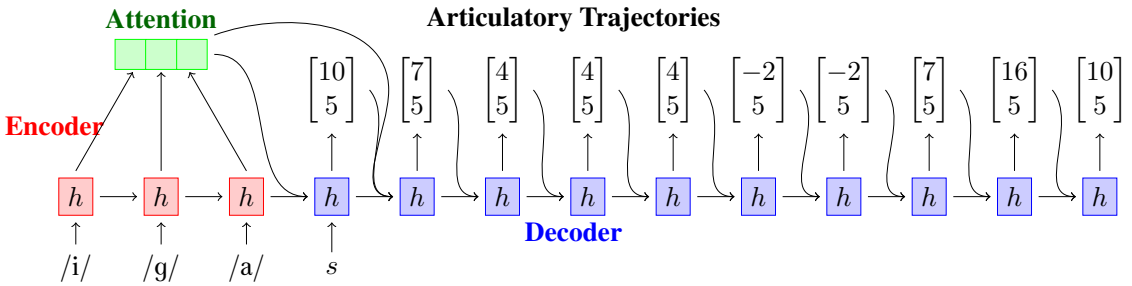

Figure 8: Sequence-to-sequence recurrent neural network with encoder-decoder attention. For sample input /ig-a/, the model outputs at each timepoint a two-dimensional vector containing predicted values for tongue body and lip constriction degrees. While attention is shown as part of the input only to the first two decoder hidden states, it is part of the input for all hidden states in our models.

The vector a containing a weighting scalar a_i for each encoder hidden state h_i is then passed through a $softmax$ function to produce a probability distribution over encoder hidden states. This distribution is then used to perform a weighted sum of all of the encoder hidden state vectors h_i in matrix H to produce the attention vector w_t, as in Equation 4.

$$w_t = softmax(a)H \tag{4}$$

The vector w_t containing a weighted sum of encoder hidden states is then concatenated with the decoder output of the previous timestep, $\hat{y}_{t-1}$. The resulting vector is the decoder's x_t, which is then input to the hidden layer along with the hidden layer from the previous time step, h_{t-1}, as in Equation 5.

$$decoder\ x_t = concat(\hat{y}_{t-1}, w_t) \tag{5}$$

Finally, the decoder produces a two-dimensional output vector $\hat{y}_t$ by multiplying the hidden state vector h_t by its weight matrix W_o, as in Equation 6. One value of $\hat{y}_t$ corresponds to predicted lip aperture, and the other to predicted tongue body height.

$$\hat{y}_t = W_o h_t \tag{6}$$

At the first timepoint of the decoder, there is no previous output $\hat{y}_{t-1}$ to concatenate with w_t as in Equation 5. In many encoder-decoder implementations, the first input to the decoder is the embedding for a special start-of-sequence token. However, in our model the decoder inputs and outputs are not embeddings of words from a fixed vocabulary, but rather continuous values for lip and tongue body constriction degree. Because the use of a special start-of-sentence token was not available to us, we instead implemented the first input to the decoder as a two-dimensional vector s whose values were learned by the model throughout training.

The full sequence-to-sequence model architecture is illustrated in Figure 8.

3.3 Training

We trained twenty models on data conforming to the Nzebi-like stepwise height harmony pattern described in Section 3.1.

Loss for a given trial was computed as the sum of the squared error, summed across all ten output timepoints and both articulators (lips and tongue body), as in Equation 7.

$$L(\hat{y}, y) = \sum_{art=1}^{2} \sum_{t=1}^{10} (y - \hat{y})^2 \tag{7}$$

This loss was back-propagated through the model after each trial. We used the Adam optimizer (Kingma and Ba, 2015) and a learning rate of 0.001 to perform model parameter updates. We trained each model for 200 epochs, at which point improvement on loss appeared to plateau for all models.

Because the primary focus of the current work is on model interpretability rather than model performance, we made the decision not to partition the data into training and test sets.

4 Results and Discussion

4.1 Model Performance

All models were able to learn to produce the training data with a high degree of accuracy. Across the twenty models, the mean loss per word after 200 epochs was 3.51.

To illustrate model performance, we provide one model's output articulatory trajectories for the lips and tongue body for the input forms /eb-a/, which

should be produced faithfully as [eba], and /eb-i/, which should be produced as [ibi] due to height harmony. As seen in figures 9 and 10, the predicted output trajectories for both vocal tract articulators closely match the target trajectories. Importantly, the model has learned that the /e/ of root /eb/ should be produced with a constriction degree of 8 (narrow-mid) before nonhigh vowels and a constriction degree of 4 (narrow) before high vowels that trigger harmony, correctly producing the Nzebi-like stepwise vowel raising pattern.

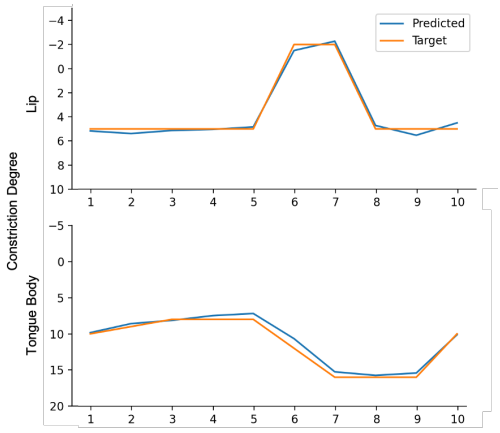

Figure 9: Input /eb-a/ correctly produced as [eba]

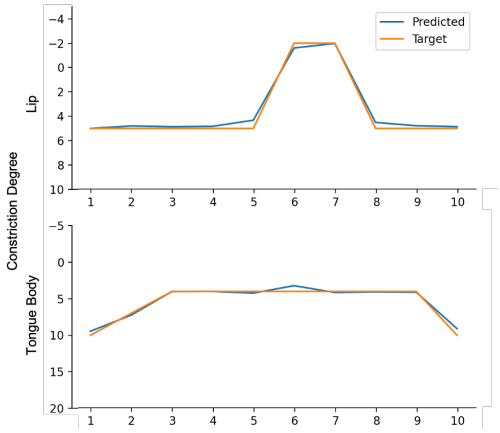

Figure 10: Input /eb-i/ correctly produced as [ibi]

4.2 Model Interpretation

In order to probe our models for linguistic structure, we examined patterns of attention between encoder and decoder hidden states. We hypothesized that the patterns of gestural activation in the gestural score for a given form could be reflected in patterns of encoder-decoder attention. For instance, if the decoder at a certain timepoint attended highly to an encoder hidden state associated with a certain input segment, it could be interpreted as the gesture associated with that input segment being active (i.e. affecting the state of the vocal tract) at that timepoint. In this way, we could use these encoder-decoder attention maps as an analog of the gestural score for a given word.

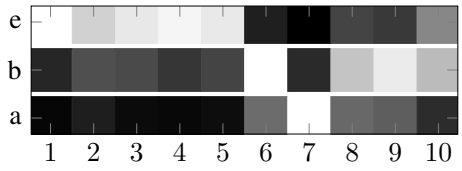

Figure 11: **Non-harmonizing Form**: Attention over input segments (vertical) at each decoder timepoint (horizontal) for input /eb-a/. Lighter squares represent more attention on that segment.

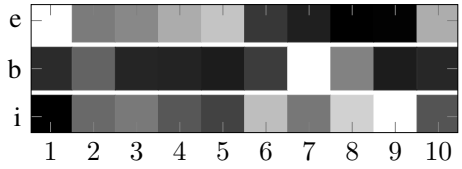

Figure 12: **Harmonizing Form:** Attention over input segments (vertical) at each decoder timepoint (horizontal) for input /eb-i/. Lighter squares represent more attention on that segment.

For this analysis, at each timepoint of the decoder we recorded the softmaxed vector of attention weights that determine how much or how little each encoder hidden state affects the decoder hidden state. To control for cases in which an attention weight could be artificially inflated or deflated according to the magnitudes of the values in its associated encoder hidden state vector, each attention weight was normalized by multiplying it by the magnitude of that vector.

Figures 11 and 12 show the attention heatmaps for items /eb-a/ → [eba] and /eb-i/ → [ibi], whose outputs were shown in figures 9 and 10. In these heatmaps, high attention paid to an input segment at a certain timepoint is indicated by a lighter square, while low attention is indicated by a darker square.

The word /eb-a/ → [eba] contains no harmony trigger, and we would therefore expect no vowel overlap, but rather activation of the gesture of [e] followed by activation of the gesture of [a], as in the gestural score for a VCV sequence in Figure 2. The attention map in Figure 11 shows that input segment /e/ is highly attended to during the first five decoder timepoints, while input segment /a/ is

attended to during the last five decoder timepoints. The medial input consonant /b/ is highly attended to at timepoint 6, corresponding to the timepoint at which the target constriction degree for the consonant is first achieved. All of these patterns of encoder-decoder attention are consistent with patterns of gestural activation we would expect to see in the gestural score for a non-harmonizing VCV sequence.

In the word /eb-i/ → [ibi], the second vowel is a harmony-triggering high vowel, and the first vowel raises one step along the height scale relative to its intrinsic target tongue body constriction degree. According to Smith's (2020b) analysis of Nzebi height harmony, this raising is the result of overlap of the first vowel gesture by the second, as in Figure 4. The attention heatmap in Figure 12 is consistent with this analysis. Again, input segment /e/ is attended to during the first five decoder timepoints, and input consonant /b/ is highly attended to at a timepoint during which is has achieved its target constriction degree. However, for this word the input segment /i/ is attended to, at least to some degree, during all decoder timepoints associated with active (i.e. non-neutral) positioning of vocal tract articulators. This suggests that harmony-triggering /i/ affects the state of the vocal tract throughout the production of the word [ibi], and not just during the production of the second syllable. We interpret this as a result consistent with the analysis of Nzebi height harmony within the Gestural Harmony Model: harmony is the result of the extended activation of a harmony-triggering gesture, such that that gesture overlaps the gestures of surrounding segments, the undergoers of harmony.

Variable	β	SE	P($>\|t\|$)	
(Intercept)	1.64	0.12	< 0.001	***
high V_2	0.28	0.024	< 0.001	***
time	0.005	0.010	0.61	

Table 3: Summary of fixed effects for linear mixed effects model with attention on V_2 as a dependent variable, and random intercepts by model.

To test whether this result held widely among the forms produced by all of our models, we ran a linear mixed effects model over all two-syllable (V_1CV_2) sequences produced by our twenty models using the lme4 package in R (Bates et al., 2015). This analysis focused on the first five time steps of the decoder, when V_1 is produced, either blended or unblended with V_2. We used the identity of input V_2 as either a high harmony trigger (/i, u/) or a non-high non-trigger and decoder timepoint as main factors, model as a random factor, and the attention value assigned to the encoder hidden state associated with input V_2 as the dependent variable. We found a significant effect of input V_2 identity (whether V_2 was a high vs. non-high vowel) on the attention paid to input V_2's associated hidden state ($p < 0.001$; further summary statistics can be found in Table 3). This result suggests that the decoder learns to pay more attention to a V_2 at an earlier timepoint when that V_2 is a harmony trigger, consistent with the representation of an anticipatory (early-activating) gesture assumed by the Gestural Harmony Model.

5 Conclusion and Future Work

In this paper, we have shown that sequence-to-sequence neural network models of speech production with encoder-decoder attention develop emergent structure analogous to the symbolic representations of the Gestural Harmony Model. In particular, we have shown that our GestNet models attend to their encoder hidden states in a pattern similar to the timecourses of gestural activation represented in a gestural score. We show that unlike most vowels, harmony triggers are attended to throughout the decoder's outputs of articulatory trajectories, mirroring the analysis of triggers of regressive harmony as anticipatory gestures in the Gestural Harmony Model.

While our current work has successfully found evidence for the extended activation of harmony-triggering gestures, there are still many avenues for the development of additional methods for model interpretability. This paper has focused on whether encoder-decoder attention maps captured the anticipatory nature of harmony triggering suffix vowels in VCV sequences. Future work should investigate how well attention maps perform at matching assumed gestural scores for longer words and larger and more complex lexicons.

Another open question involves whether we can expect to find emergent linguistic structure corresponding to additional gestural parameters in GestNet and other neural network models of speech production. Future work should probe our neural models for evidence of these parameters; for instance, gestural strength might be analogous to the magnitude of the hidden state vector that is

weighted by the encoder-decoder attention mechanism.

Another potentially fruitful domain of investigation is intergestural coordination. An intriguing question is whether gestures in particular coordination relations result in particular patterns of attention during one another's production. Deeper investigation of attention maps in GestNet models may also further shed light on which types of phonological processes are better understood as resulting from gestural overlap and/or re-coordination, and which are better understood as alternations based on changes in the gestural makeup of phonological forms.

Acknowledgements

For helpful discussion of this work, we thank the members of the Neurosymbolic Computation Lab at JHU: Coleman Haley, Najoung Kim, Matthias Lalisse, Tom McCoy, and Paul Soulos. This work is supported by Microsoft Research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.

Catherine P. Browman and Louis Goldstein. 1986. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252.

Catherine P. Browman and Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology*, 6(2):201–251.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

Carol A. Fowler and Elliot Saltzman. 1993. Coordination and coarticulation in speech production. *Language and Speech*, 36:171–195.

M. Gareth Gaskell, Mary Hare, and William D. Marslen-Wilson. 1995. A connectionist model of phonological representation in speech perception. *Cognitive Science*, 19(4):407–439.

Malcolm Guthrie. 1968. Notes on Nzɛbi (Gabon). *Journal of African Languages*, 7(2):101–129.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference for Learning Representations*.

Robert Kirchner. 1996. Synchronic chain shifts in Optimality Theory. *Linguistic Inquiry*, 27(2):341–350.

Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. In *Proceedings of the Society for Computation in Linguistics: Vol. 3*.

Nicole Mirea and Klinton Bicknell. 2019. Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1595–1605.

Hosung Nam, Louis Goldstein, Elliot Saltzman, and Dani Byrd. 2004. TADA: An enhanced, portable Task Dynamics model in MATLAB. *Journal of the Acoustical Society of America*, 115:2430.

Frederick Parkinson. 1996. *The Representation of Vowel Height in Phonology*. Ph.D. thesis, The Ohio State University.

Brandon Prickett. 2019. Learning biases in opaque interactions. *Phonology*, 36(4):627–653.

Jennifer Rodd. 1997. Recurrent Neural-Network Learning of Phonological Regularities in Turkish. In *CoNLL97: Computational Natural Language Learning, ACL*, pages 97–106.

Eric Rosen. 2021. Lexical strata and phonotactic perplexity minimization. In *Proceedings of the Society for Computation in Linguistics: Vol. 4*.

Elliot Saltzman and Kevin G. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382.

Miikka Silfverberg, Lingshuang Jack Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics: Vol. 1*, pages 136–144.

Caitlin Smith. 2016. A gestural account of neutral segment asymmetries in harmony. In *Proceedings of the 2015 Annual Meeting on Phonology*.

Caitlin Smith. 2018. *Harmony in Gestural Phonology*. Ph.D. thesis, University of Southern California.

Caitlin Smith. 2020a. Partial height harmony as partial transparency. In *Proceedings of the 2019 Annual Meeting on Phonology*.

Caitlin Smith. 2020b. Stepwise height harmony as partial transparency. In *Proceedings of the 50th Annual Meeting of the North East Linguistic Society*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 3104–3112.

Sam Tilsen. 2020. A different view of gestural activation: learning gestural parameters and activation with an RNN. Manuscript, Cornell University.