Gestural Scores in Recurrent Neural Networks Trained on Two-Dimensional Vowel Harmony

Hilly Steinmetz
University of Washington

1 Introduction

Recent work has shown that the Gestural Harmony Model–a modified version of articulatory phonology–can capture and explain several phenomena in vowel harmony, including blocking, transparency, and bidirectional harmony. In her dissertation, Smith argues that left-to-right harmony can be explained by an additional gestural parameter: persistence (Smith, 2018). Normally, gestures are self-deactivate–the articulators return to their baseline state–but harmony occurs when a gesture continues to be maintained over the course of a word.

Smith et al. (2021) found additional support for the Gestural Harmony Model with recurrent neural networks (RNNs). They trained RNN encoder-decoder models with attention on a constructed dataset based on Nzebi height harmony. They found that the networks pay greater attention to harmonizing segments than non-harmonizing segments. While neural networks are not analogous to human cognition, it does support provide support for the idea that gestural targets can be influenced by other segments.

This paper follows up on Smith et al. (2021) research by training the same model on a two-dimensional harmony system based on Turkish. I find the addition of another harmony system causes models to pay more attention to harmony-triggering segments. I also find that rounding significantly contributes to the attention models pay to the harmony-triggering segment. I argue that these results can be explained by the fact that rounding harmony is less predictable than backness harmony.

In section 2, I review Turkish vowel harmony and the Gestural Harmony Model. Section 3, outlines the questions that motivated this study. Section 3 provides an overview of encoder-decoder RNNs, and how they've been used in phonology studies. Section 5 details how I constructed the data and modified the model from Smith et al. (2021) to learn a 2-dimensional vowel harmony system. Section 6 describes the outcomes of my study. Section 7 and section discuss the results and the implications of my findings.

2 Vowel harmony and the Gestural Harmony Model

2.1 *Gestural Harmony Model* Vowel harmony is a long-distance phonological process whereby vowels trigger alternations in adjacent vowels, even when there are intervening consonants. In general, the vowels end up taking on features more similar to the vowel that triggered the harmony. Vowel harmony is highly sensitive to phonological knowledge–the segments that undergo harmony, are transparent to harmony, or block harmony, are not easily defined by natural classes (Krämer, 2003). For instance, Halh Mongolian has rounding harmony and pharyngeal harmony whereby vowels can be divided into the following classes (Svantesson et al. 2005 as cited in Smith):

	non-pharyngeal		**pharyngeal**	
	non-round	round	non-round	round
high	i	u		ʊ
non-high	e	o	a	ɔ

Table (1) shows examples of vowel alternation patterns in Halh Mongolian. Normally, the initial syllable in Mongolian will trigger rounding harmony as in (a-c). However, [i] is transparent to vowel harmony (d-f) even though [u] and [ʊ] are not transparent when the causitive past affix is included.

	a. [og-lʒo] ‘give (past)’	d. [poːr-ig-o] ‘kidney (acc. refl.)’	g. [oɡ-ulʒ-lʒe] ‘go (caus. past)’
(1)	b. [ɔr-lʒɔ] ‘enter (past)’	e. [xɔːlʒ-iɡ-ɔ] ‘food (acc. refl.)’	h. [ɔr-ʊlʒ-lʒa] ‘enter (caus. past)’
	c. [it-lʒe] ‘eat (past)’	f. [piːr-iɡ-e] ‘brush (acc. refl.)’	i. [it-ulʒ-lʒe] ‘eat (caus. past)’

Table 1: Mongolian vowel harmony and blocking

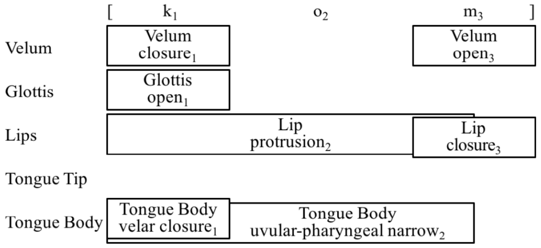
(Svantesson et al., 2005)

In addition, vowel harmony is often tied to lexicon. In Turkish, some suffixes from foreign loan words don’t harmonize with the root when the accusative suffix /-I/[1] is added, as in [sat], [sat-i] ‘watch’ and [rol], [rol-y] ‘role’ (Schaaik, 2020). Still, loanwords with high vowels that require epenthesis to meet Turkish phonotactic requirements do show vowel harmony as in [fylyt] ‘flute’. Those with non-high vowels, like the borrowing for ‘flirt’, show more alternation in the vowel epenthesized between the segments [f] and [l] ([filœrt ∼ fylœrt]) (Kaun, 2004). The typological and lexical variation make vowel harmony systems difficult to capture for many phonological frameworks.

Previous attempts to analyze vowel harmony used autosegmental representations of phonology. Kaun, for example, uses autosegmental representations alongside optimality theory to explain rounding harmony. She posits a constraint GESTUNI[(ROUND)], along with several alignment constraints, require all linked autosegments must have the same rounding articulation. She uses these constraints to create a typology of rounding harmony across several languages and language families(Kaun, 2004). Others have proposed more complex sets of constraints to prevent all autosegments from linking to a single feature. (Krämer, 2003) uses correspondence theory to posit constraints that link each subsequent vowel to one another but prevent all the segments from being in correspondence with each other. So, in a word $CV_aCV_bCV_c$, V_a and V_b are in a correspondence relationship, V_b and V_c are in a correspondence relationship, but V_a and V_c are not in a correspondence relationship. Overall, it can be difficult to represent vowel harmony in many classical frameworks, and the complexities of these representations only grow as phonologists try to analyze more complicated harmony systems.

The Gestural Harmony Model can provide a more parsimonious analysis of vowel harmony. It leverages the representations of articulatory phonology (AP), which represent phonological segments as a series of articulatory targets. An advantage of AP is that gestures can overlap and blend, which can explain the occurrence of surface forms in the grammar.

Figure 1: The word “comb” represented in as a gestural score (from Smith 2018, p. 15)



Smith (2018) adds two parameters to gestures in order to create the Gestural Harmony Model: persistence and anticipation. In the case of persistence, the gesture does not self-deactivate once the articulatory target is reached, allowing it to spread to the remainder of a word segment. Smith (2018) argues that Kyrgyz, another Turkic language with rounding harmony, specifies in its grammar to not deactivate lip protrusion after certain triggering vowels. Persistence therefore can account for left-to-right harmony. Anticipation, on the other hand, accounts for right-to-left harmony. If a grammar gives weight to anticipation, then any gestures preceding a particular gestural target will also adopt the same target. The following example from Smith (2018) illustrates how the gestural representations can be used in constraint-based grammar to explain how Kyrgyz vowel harmony arises even when the underlying vowel is not specified at triggering harmony

[1] I use /I/ to denote the underspecified high vowel which can be realized as [u], [i], [y], or [ɯ]

(the notation here is imprecise, but it should provide a sufficient overview of working with the model).[2]

Input: /tu$_1$da$_2$n/ [u$_1$ a$_2$] [deactivate$_{[LP]}$]	PERSISTENCE(LP)	SELFDEACTIVATE	IDENT-IO(DEACTIVATE)
a. [tu$_1$da$_2$n] [u$_1$ a$_2$] [deactivate$_{[LP]}$] [LP No LP]	*!		
☞ b. [tu$_1$do$_2$n] [u$_1$ o$_2$] [persist$_{[LP]}$] [LP]		*	*

(Smith, 2018:p. 78)

A key advantage to the Gestural Harmony Model is that it can account for patterns like blocking and transparency in a parsimonious fashion. In the case of transparency, the articulator stays at its target positional. If intervening segments are composed of a compatible gesture, they don't cause the maintained gesture to deactivate. The other side of this claim is that segments block harmony when the gestural targets come into conflict. With these tools, the Gestural Harmony Model can explain more difficult patterns of vowel harmony like iterative harmony in Yakut (Smith, 2018) and (Smith, 2019) partial transparency.

The Gestural Harmony Model might also be empirically supported through motion-tracking studies of lip movement. Boyce (1990) found that Turkish speakers and English speakers differ in lip movement patterns when articulating words with a [u C u] pattern. Turkish speakers will keep their lips protruded throughout the utterance, producing a plateau-like pattern when the degree of protrusion is graphed. English speakers, on the other hand, will bring their lips closer to their resting state while producing the consonant, producing a trough-like pattern when the degree of protrusion is graphed. Smith (2018) argues that these findings support the Gestural Harmony Model since the model predicts ongoing and overlapping lip protrusion gestures.

2.2 *Turkish vowel harmony* Like Kyrgyz, Turkish has a two-dimensional vowel harmony consisting of palatal (backness) harmony and rounding harmony. Its vowel inventory can be categorized by size dimensions: height, backness, and roundedness (Schaaik (2020), Mielke (2011)).

	front		**back**	
	non-round	round	non-round	round
high	i	y	ɯ	u
neutral		œ		o
low	e		a	

Turkish suffixes often have vowels that are underspecified backness and roundedness. As a result, suffixes can have several alternations. For example, the accusative suffix /-I/ can surface as one of four forms depending on the verb. Similarly, the plural suffix /-lLr/ can surface as /-ler/ or /-lar/.

nom.	acc.	gloss	sg.	pl.	gloss
top	topu	'ball'	armut	armutlar	'pear'
sepet	sepeti	'basket'	ʃiʃe	ʃiʃeler	'bottle'
ak	akɯ	'egg'			
kœpry	kœpryjy	'bridge'			

(Schaaik, 2020)

As Krämer (2003) notes, in Turkish palatal and rounding harmony can occur independently of one another in high vowels. Still, it might be wrong to claim that rounding and palatal harmony are totally

[2] LP here stands for Lip Protrusion. I did not have the time to properly fit the correct notation into this paper. See Smith (2018) for the precise notation.

independent of one another. Kaun (2004) notes that this may be an exception because of a typological relationship between the two: there are very few languages that just have rounding harmony. Most languages with rounding harmony also have backness or height harmony. Moreover, rounding harmony is often highly conditioned by other phonological features like backness, height, or even the features of adjacent consonants (Kaun, 2004). Because rounding harmony in Turkish is largely unconditioned, it can serve as an interesting starting point to answer questions about articulatory, perceptual, or informational pressures on vowel harmony systems.

3 Is vowel harmony a perceptually motivated?

There is no definitive answer to the linguistic purpose of vowel harmony, especially in two-dimensional vowel harmony systems. Kaun (2004) argues that the addition of rounding increases the perceptual salience of certain segments. In other words, making adjacent segments more similar to one another increases the chances of a vowel being properly perceived. She also cites studies that found that perceptual salience can explain why rounding harmony systems tend to only apply to high vowels–the perceptual salience of rounding vowels is greatest for high vowels.

The claim that vowel harmony serves as a perceptual aid might be supported by a psycholinguistic study conducted by Beddor et al. (2013) which found that English and Thai speakers were better at judging vowels that preceded nasal consonants (as in the word ‘candle’) when the vowel was nasalized. They argue that listeners leverage contextual information to perceive vowel segments. This finding supports the claim that vowel harmony enhances the perceptual salience of vowels since, like coarticulation, vowel harmony involves the spreading of a certain feature like roundedness or height.

Still, the question remains as to whether vowel harmony serves primarily as a perceptual or articulatory aid. Additionally, if vowel harmony is primarily a perceptual aid, then vowel harmony’s associated acoustic or informational cues could both contribute to its perceptual salience. Several studies show that vowel harmony systems–even ones with difficult to predict patterns–display statistical regularities (Hayes & Londe 2006, Caplan & Kodner 2018). The broader question motivating this paper is, therefore: *can vowel harmony provide informational cues, and, if so, how*? The specific research question this paper is concerned with is *does vowel harmony propagate informational cues across a sequence in a way that facilitates predictions of subsequent sequences*? One manner to study the ways phonological information helps predict future subsequent segments is through the use of recurrent neural networks.

4 Using Recurrent neural networks to study vowel harmony

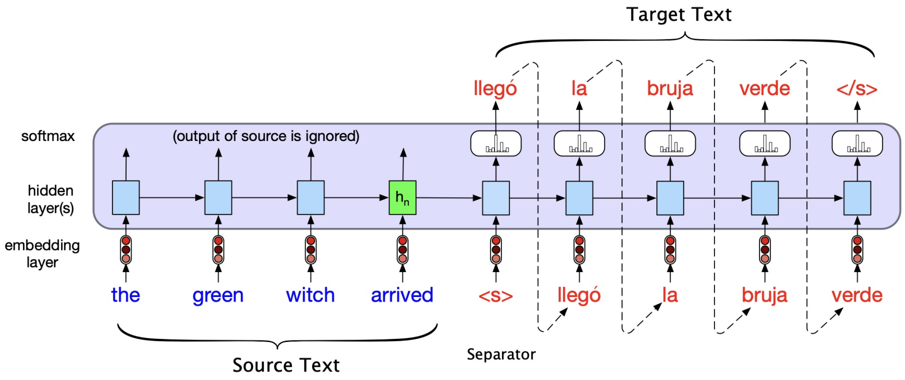
4.1 *A brief review of neural networks* Artificial neural networks are a technique used to learn statistical patterns in data. They’re loosely inspired by the way neurons work: an artificial neuron propagates information forward (in the form of matrices of numbers) to other neurons if its input meets a certain threshold. These thresholds can be learned with minimal supervision through the back-propagation algorithm (Jurafsky & Martin, 2022). Neural networks can be arranged into different architectures, many of which improve performance on particular tasks.

A recurrent neural network (RNN) is useful for processing sequences since it preserves information from the previous items in the sequence. An RNN processes items in a sequence individually. At each time step, it calculates a hidden state that it preserves and uses to calculate the output of the time step of the sequence. (Jurafsky & Martin, 2022)

RNNs can be arranged into an encoder-decoder architecture, which is particularly useful for training models to translate one sequence to another. In this architecture, one RNN encodes an entire sequence into a hidden state. This hidden state is then passed into a decoder–an RNN that predicts the target sequence. The predictions at each time step of the decoder influence the prediction of the next element in the target sequence by passing forward its hidden state.

Attention is an additional mechanism that is often added to encoder-decoder models. Because sequences are passed as inputs to RNNs one at a time, each hidden state is substantially influenced by the most recent hidden state. In practice, this means that RNNs can “forget” non-recent information (Jurafsky & Martin 2022, Smith et al. 2021). In encoder-decoder architectures, attention involves calculating the similarity (e.g., cosine similarity) between the decoder’s hidden state at the current time step and the encoder’s hidden state.

Figure 2: The architecture of an encoder-decoder RNN (from Jurafsky & Martin 2022, Ch. 10, p.8)

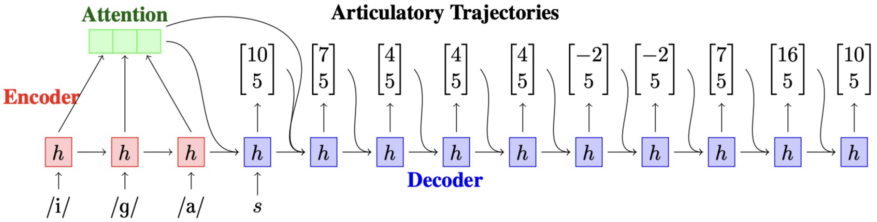


In other words, at each time step in the decoder, attention discovers what time step of the encoder is most similar to the current time step. These values are used to influence the decoder's predictions. An additional benefit of attention is that it is interpretable (Smith et al. 2021). Unlike neural networks where behavior is emergent, attention provides a clear relationship that can help researchers identify what correspondences the model considers to be the most important.

Encoder-decoder models with attention have been studied by phonologists to conduct research, oftentimes to study open questions in cognitive science. For instance, Kirov & Cotterell (2018) was able to show that RNNs trained on the English past tense data, were successfully able to generalize their predictions of regular and irregular past tense to unseen data. They found that the models predicted the past tense of 'rife' as 'rifed' or 'rofe' with probabilities that resembled the distribution of human responses. They also found that the learning pattern resembled early childhood learning. A rejected anonymous submission to ACL 2022 trained small encoder-decoder RNNs with attention to Turkish and Finnish data. They found that the representations of vowels learned by the decoder clustered front and back vowels together, indicating that it might learn to represent vowel harmony ("Phonological learning and encoding with small RNNs", 2022).

4.2 *Smith et al. (2021): RNNs and emergent gestures* Smith et al. (2021) sought to train RNNs to map a sequence of phonemes to gestural targets. The authors were interested in whether neural networks could learn representations consistent with the Gestural Harmony Model. They trained an encoder-decoder RNN with attention (see figure 3) to learn gestural targets for vowel height and lip closure from CV and VCV sequences. They used a toy dataset that was constructed to resemble Nzebi, which has height harmony. In Nzebi, vowels are raised by one step ([a] → [ɛ], [ɛ] → [e], [e] → [i]) in when it is followed by a high vowel. The gestural targets they used were determined after synthesizing speech in TADA (Hosung Nam & Byrd, 2004). The segments and gestural targets used to construct the data are shown in **??**.

Figure 3: The architecture of encoder-decoder with attention used by Smith et al. (2021) (from Smith et al. 2021, p. 66)


Attention
Articulatory Trajectories
Encoder
$\begin{bmatrix}10\\5\end{bmatrix}$ $\begin{bmatrix}7\\5\end{bmatrix}$ $\begin{bmatrix}4\\5\end{bmatrix}$ $\begin{bmatrix}4\\5\end{bmatrix}$ $\begin{bmatrix}4\\5\end{bmatrix}$ $\begin{bmatrix}-2\\5\end{bmatrix}$ $\begin{bmatrix}-2\\5\end{bmatrix}$ $\begin{bmatrix}7\\5\end{bmatrix}$ $\begin{bmatrix}16\\5\end{bmatrix}$ $\begin{bmatrix}10\\5\end{bmatrix}$
h
Decoder
/i/
/g/
/a/
s

They trained 20 models using this data and then analyzed the attention the models paid to the following vowels when generating outputs gestural targets. They found that models paid significantly more attention to the final vowel in VCV sequences when it was a height harmony triggering vowel. They interpreted these results as indicating that analogs to gestural scores (here represented by attention) emerge in neural networks. In other words, gestural targets are maintained by RNNs through the attention mechanism.

5 Comparing RNN models trained on different harmony systems

The present study investigates whether the addition of a second harmony system increases or decreases attention paid to the harmony-triggering vowel. Because the hidden state also contributes to predictions made by the model, it is not clear whether an encoder-decoder RNN would pay additional attention to harmony-triggering vowels when another harmony system is added. In fact, it might decrease attention since it provides the model with additional contextual clues. Therefore, I proposed the following hypotheses for the model's performance in a two-dimensional harmony system:

H1 Attention increases because the model needs to produce two an additional gestural target based on the harmony-triggering vowel.

H2 Attention decreases because more information can be encoded into the hidden state.

These hypotheses would reveal important aspects of how the information is propagated by vowel harmony systems. If **H1** proves to be correct, then vowel harmony might increase the salience of a segment (with implications for segment perception in human language). **H2**, on the other hand, is more analogous to coarticulation–the additional harmony system helps hidden layers propagate information forward, meaning that there is less need for attention to reinforce its salience. In other words, **H2** might imply that the two systems can conspire to reduce the attention needed to pay to earlier segments. This would be analogous to the claims made by Kaun (2004) and Beddor et al. (2013) that coarticulation makes the surrounding segments more salient.

5.1 *Data* I constructed a toy dataset inspired by Turkish vowel harmony. I constructed the data by interpolating from the dataset constructed by Smith et al. (2021). The inputs consist of CV and VCV segments and the outputs correspond to gestural targets. The consonants consists of [d], [b], and [g]. I added the consonant [d] to have sufficient data since a dataset with just [b] and [g] might be too small. I chose [d] also because its gestural targets did not overlap with any of the vowel's targets. The suffix vowels were represented by [L] or [H], where [L] is a low vowel underspecified for backness and [H] is a high vowel underspecified for backness and rounding.

The values of these gestural targets are derived from the software TADA (Hosung Nam & Byrd, 2004). I used the software to ensure that the speech synthesized by the software roughly matched the phonemic transcription and that the targets could generalize to Turkish data. The data consists of three articulators: lip aperture (LA), tongue body closure degree (TBCD), and tongue body constriction location (TBCL). LA denotes how constricted the lips are from -2 (closed) to 11 (open). An LA of 5 is used for rounded vowels. TBCD roughly corresponds to vowel height where a lower number is a higher vowel. TBCL roughly corresponds to vowel backness, where larger values denote greater backness.

Table 2: The segments and target used to train RNN in the present study.

	Articulatory Targets		
Segment	**LA**	**TBCD**	**TBCL**
i	11	4	95
u	5	4	150
ɯ	11	4	150
y	5	4	95
œ	5	8	95
o	5	8	150
e	11	8	95
a	11	16	150
d	NA	NA	NA
b	-2	NA	NA
g	NA	-2	110

LA is Lip Aperture, TBCD is Tongue Body Constriction Degree and TBCL is Tongue Body Constriction Location.

I made several simplifying assumptions while constructing the data. One assumption was that the gestural targets for Nzebi phonemes were exactly the same as those for my Turkish phonemes. I also assumed that high, medium. and low vowels had the same tongue body closure targets, and that front and back vowels had the same tongue body constriction targets, even though this may not be reflective of Turkish vowels.[3] The segments are gestural targets are shown in Table 2.

5.2 *Adapted model* The model largely resembled the one used in Smith et al. (2021). The code for their model is publicly available on Github [4]. I preserved most of the original code but made some modifications for the present study so that it be adapted to a two-dimensional harmony system.[5]

Because I wanted to compare how the model responded to both two-dimensional and one-dimensional harmony, the losses for each articulatory target were averaged instead of summed. Additionally, because the values for velar location were much larger than lip aperture and tongue body closure, I normalized the values so that they fall between [-1, 1] by scaling each series by its range. This prevents any one articulator from influencing the training algorithm more than another. I also wanted to prevent the CV sequences—whose values at time steps 6 to 10 are 0 to represent a resting articulator—from influencing model training, so I masked the losses for CV segments at time steps 6 to 10 by setting the losses to zero. In other words, the model will not be concerned with modelling these time steps.

5.3 *Experiment procedure* I created 3 conditions, each with a different harmony system:

Condition 1 (C1): Train models with just backness harmony. Predict outputs for TBCD and TBCL.

Condition 2 (C2): Train models with rounding harmony *and* backness harmony. Predict outputs for LA, TBCD, and TBCL.

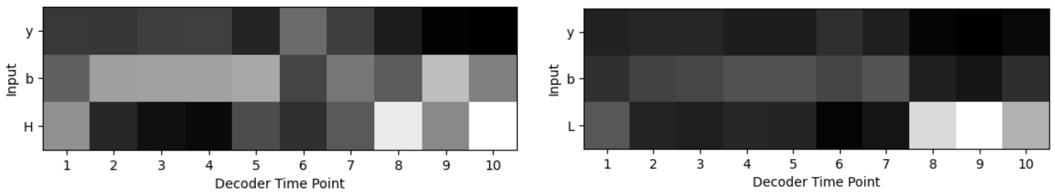
Condition 3 (C3): Train models with just rounding harmony. Predict outputs for LA and TBCD.

For each of the 3 conditions, I trained 20. Each model was over 200 epochs, with a learning rate of 0.00075 and using the PyTorch package's AdamW optimizer (Paszke et al., 2019). The loss values between the predicted and target outputs were calculated with mean squared error, and average over time steps and articulators. Model attention and losses at each time step were calculated for every VCV sequence. The results were analyzed in R.

6 Results

Each condition produced models that were able to predict articulatory targets at each time step with a high degree of accuracy. The conditions produced models that differed in attention and accuracy.

Figure 4: Attention scores for the input /yb-H/ and /yb-L/ in model gestnet_la_tb_tc_8_200.pt. This model was trained to predict targets for all three articulators. Lighter colors indicate higher attention.



[3] The values for TBCD and the resting state for LA were constructed from a slide deck tutorial on TADA (Goldstein 2015, URL: sail.usc.edu/∼lgoldste/Ling285/Slides/Lect15_handout.pdf) and the TADA manual (Nam & Goldstein 2007, URL: sail.usc.edu/∼lgoldste/ArtPhon/TADA%20stuff/TADA_manual_v09.pdf).

[4] Code for the model from Smith et al. (2021): https://github.com/caitlinsmith14/gestnet.

[5] Code for the adapted model can be found on Github: https://github.com/hasteinmetz/phonology-paper.

6.1 *Attention* My analysis found that C2 and C3, the conditions with just rounding harmony and two-dimensional harmony, paid significantly more attention to the first vowel between time steps 5 and 10. However, C1 did not pay significantly more attention. In fact, the models in C2 on average did not pay much attention to the first segment compared to the other groups.

Table 3: Mean attention scores for the last five time steps by condition.

Condition	V_2	Mean	Standard deviation
Backness only	High	0.157648	0.212985
	Low	0.183318	0.244486
Rounding only	High	0.409541	0.321467
	Low	0.397846	0.341531
Rounding and backness	High	0.439764	0.353135
	Low	0.464506	0.358871

To confirm this observation, for each of the conditions, I ran a mixed-effects linear model using the `lme4` package in R (Bates et al., 2015). I used attention in the last 5 time steps as the dependent variable, V_2 and time step as independent variables, and the model as a random variable. This was the analysis performed by Smith et al. (2021).

Table 4: Summary of linear effects models for each condition.

Condition	Variable	β	SE	$P(> \|t\|)$	
Backness only	(Intercept)	1.320e-01	2.052e-02	2.12e-08	***
	Low V_2	-3.849e-04	1.916e-03	0.752	
	Time	-3.849e-04	1.916e-03	0.841	
Rounding only	(Intercept)	3.185e-01	2.644e-02	$<$ 2e-16	***
	Low V_2	-2.115e-02	7.604e-03	0.00543	**
	Time	-7.034e-03	2.688e-03	0.00891	**
Rounding and backness	(Intercept)	-2.413e-01	2.782e-02	1.73e-12	***
	Low V_2	6.861e-02	2.650e-03	$<$ 2e-16	***
	Time	7.158e-02	7.495e-03	$<$ 2e-16	***

To compare the conditions, I calculated the average the attention scores across models, reducing the number of observations to 720. I then used these scores to conduct a two-way ANOVA with time and condition (C1, C2, C3) as independent variables and attention as the dependent variable. I found condition to play a significant effect in attention scores ($F = 65.72$, $Pr(> F) < 2e{-}16$). Tukey's honest significance test found significant differences between condition 2 and condition 1 ($p = 0, CI = [0.12859933, 1.992914e - 01]$) and condition 3 and condition 1 ($p = 0, CI = [0.09320032, 1.638923e - 01]$). The test also showed that the differences between attention in the C2 (rounding only) and C3 (two-dimensional) to be marginally significant ($p = 0.0495587, CI = [-0.07074503, -5.300439e - 05]$).

6.2 *Accuracy* I also wanted to see if the conditions had any effect on the performance of the models. A similar pattern emerges between conditions in accuracy measures: C1 has the better accuracy scores (lower MSE), while C2 and C3 have worse accuracy scores.

Table 5: Mean accuracy scores (mean squared error) by condition for the last five time steps.

Condition	V_2	Mean	Standard deviation
Backness only	High	0.003004	0.003560
	Low	0.003195	0.004070
Rounding only	High	0.005298	0.005443
	Low	0.007190	0.008073
Rounding and backness	High	0.003573	0.005206
	Low	0.007477	0.011945

I calculated the average accuracy scores across models for a total of 720 observations and used these values as a dependent variable in a two-way ANOVA. The ANOVA used condition and time as independent variables. The ANOVA revealed that the condition played a significant effect on accuracy scores ($F = 28.746$, $Pr(> F) < 9.81e - 13$). Tukey's honest significance test found the differences to lie between conditions 1 (fronting) and conditions 2 and 3(rounding and two-dimensional).

Table 6: Tukey HSD results on mean accuracy scores.

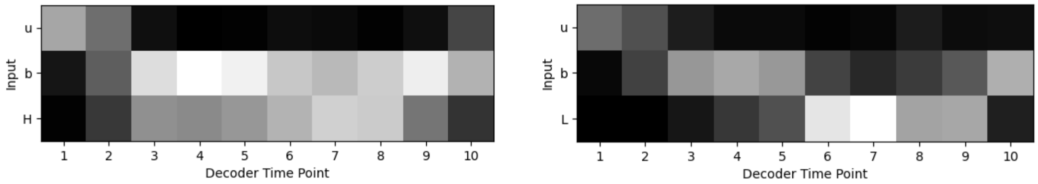
Condition Comparison	Confidence interval	p
Rounding and backness — Backness only	[0.002123604, 0.0041648476]	0.0000000
Rounding only — Backness only	[0.001404919, 0.0034461624]	0.0000001
Rounding only — Rounding and backness	[-0.001739307, 0.0003019365]	0.2239412

7 Discussion

Before discussing these findings, it is important to list some caveats about the study design. Firstly, the data may not accurately reflect two-dimensional vowel harmony systems like Turkish. Most glaringly, articulatory targets are extrapolated from Nzebi—more accurate information on how Turkish speakers produce vowels could impact these findings. Secondly, I did not perform sufficient hyperparameter tuning. I believe that the models could be improved by selecting better learning rates and investigating whether other loss functions might improve the accuracy of the models. Finally, it might be inappropriate to compare accuracies and losses across conditions since the models in each condition were trained to predict different articulatory target sequences. For instance, it may be that the gestural targets of one articulator are more difficult than another.

With those caveats in mind, we might be able to make some humble claims about the patterns observed in this study. **H1** seems to be the correct hypothesis: the addition of another harmony system *increases* the attention a model pays to the first vowel in a VCV sequence. However, the difference between two-dimensional harmony models and rounding-only harmony models was just barely significant ($p \approx 5$). It may be worth training and adding new models to these conditions to see if these rounding-only and two-dimensional harmony systems continue to differ.

Figure 5: Attention scores for the input /ub-H/ and /ub-L/ in model gestnet_tb_tc_4_200.pt. This model was trained to predict targets for just TBCD and TBCL. Lighter colors indicate higher attention.



An interesting finding may be that the backness-only models did not pay much attention at all to the first vowel of the input sequences. This could be because it is easier to encode backness in the hidden state. It might be easier to encode backness in particular because none of the data have examples where a back vowel follows a front vowel (or vice versa). The model can simply learn the gestural targets of front and back vowels and then propagate that information forward with few errors. This conjecture might also explain why the backness-only models are more accurate as well. A future study might want to choose another harmony typology that has backness harmonizes in particular contexts instead of the Turkish-inspired grammar which always exhibits backness harmony.

If this conjecture proves correct and two-way harmony models still have higher attention than rounding-only models, it would lead to some interesting questions. In other words, if we found evidence that harmony systems "conspire" to increase attention in RNNs, it might reveal some interesting characteristics about how RNNs process information, and open up other interesting questions about possible analogs in human speech processing.

7.1 *Future Questions* One way to follow up on this research is to investigate what the decoder's hidden states are doing to produce gestural targets. This question was already raised by Smith et al. (2021), and could provide evidence to my conjecture that backness-only models encode backness in the hidden state. One simple way to look into this question is to use dimensionality reduction techniques to visualize the hidden states in two- or three-dimensional space (Garcia et al., 2021).[6]

Another important research question posed by Smith et al. (2021) is how models attend to vowel harmony in longer sequences and real lexicons. It might be especially interesting to see test whether, in longer sequences, the models continue to attend to the harmony-triggering vowel or whether they attend to the most recent harmonizing vowel. It will also be interesting to see whether, like in Kirov & Cotterell (2018), these models can correctly generalize vowel harmony to unseen words and compare these results to human judgments.

8 Conclusion

In this paper, I discuss how RNNs can be used to investigate the Gestural Harmony Model. I provide an overview of how the Gestural Harmony Model analyzes vowel harmony and neural networks. I discuss how Smith et al. (2021) found patterns similar to emergent gestural scores in RNNs. I detail how I adapted the data and models from Smith et al. (2021) to design my study.

My study found that RNNs trained on two-way harmony systems pay greater attention to harmony-triggering segments than models trained on backness harmony or rounding harmony alone. Models trained on backness harmony paid barely any attention to the harmony-triggering segment, which might indicate that information on backness is propagated by the hidden state. This paper can hopefully serve as a starting point for future investigations into how analogs to the Gestural Harmony Model can be found in neural networks.

[6] This was something I was hoping to complete in time for this paper, but I was unable to validate whether the visualization I produced made sense.

References

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

Beddor, Patrice Speeter, Kevin B. McGowan, Julie E. Boland, Andries W. Coetzee & Anthony Brasher (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America* 133:4, 2350–2366, URL `https://doi.org/10.1121/1.4794366`.

Boyce, Suzanne E. (1990). Coarticulatory organization for lip rounding in turkish and english. *The Journal of the Acoustical Society of America* 88:6, 2584–2595.

Caplan, Spencer & Jordan Kodner (2018). The acquisition of vowel harmony from simple local statistics. Chuck Kalish, Jerry Zhu, Martina Rau & Timothy Rogers (eds.), *CogSci 2018*, 1440–1445.

Garcia, Rafael, Tanja Munz & Daniel Weiskopf (2021). Visual analytics tool for the interpretation of hidden states in recurrent neural networks. *Visual Computing for Industry, Biomedicine, and Art* 4:1, p. 24, URL `https://doi.org/10.1186/s42492-021-00090-0`.

Goldstein, Louis (2015). Lecture 15: Speech synthesis using tada.

Hansson, Gunnar (2020). Consonant harmony.

Hayes, Bruce & Zsuzsa Cziráky Londe (2006). Stochastic phonological knowledge: The case of hungarian vowel harmony. *Phonology* 23:1, 59–104, URL `http://www.jstor.org/stable/4420265`.

Hosung Nam, Elliot Saltzman, Louis Goldstein & Dani Byrd (2004). Tada: An enhanced, portable task dynamics model in matlab. *Journal of the Acoustical Society of America* 115:5, 2430–2430.

Jurafsky, Dan & James H. Martin (2022). *peech and Language Processing*. Online, 3rd edition (draft) edn., URL `{https://}web.stanford.edu/\~jurafsky/slp3/ed3book\_jan122022.pdf`. [Accessed June-2022].

Kager, Rene (1999). *Optimality Theory*. Cambridge Textbooks in Linguistics, Cambridge University Press.

Kaun, Abigail R. (2004). *The typology of rounding harmony*, Cambridge University Press, p. 87–116.

Kirov, Christo & Ryan Cotterell (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics* 6, 651–665, URL `https://aclanthology.org/Q18-1045`.

Krämer, Martin (2003). *Vowel Harmony and Correspondence Theory*. De Gruyter, Inc., Berlin/Boston, Germany.

Mielke, Jeff (2011). *Distinctive Features*, John Wiley and Sons, Ltd, chap. 17, 1–25.

Nam, Hosung & Louis Goldstein (2007). Tada (task dynamics application) manual.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai & Soumith Chintala (2019). Pytorch: An imperative style, high-performance deep learning library. Wallach, H., H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 8024–8035.

Schaaik, Gerjan van (2020). *The Oxford Turkish grammar*. Oxford linguistics, Oxford University Press, Oxford ; New York, NY, first edition. edn.

Smith, Caitlin (). A gestural account of neutral segment asymmetries in harmony. *Proceedings of the Annual Meetings on Phonology*, vol. 3.

Smith, Caitlin (2019). Stepwise height harmony as partial transparency. *Proceedings of the 50th Annual Meeting of the North East Linguistic Society*, vol. 3, 131–154.

Smith, Caitlin M. (2018). *Harmony in Gestural Phonology*. Ph.D. thesis, University of Southern California. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated

- 2021-11-18.

Smith, Caitlin, Charlie O'Hara, Eric Rosen & Paul Smolensky (2021). Emergent gestural scores in a recurrent neural network model of vowel harmony. *Proceedings of the Society for Computation in Linguistics*, vol. 4.

Svantesson, Jan-Olof, Anna Tsendina, Anastasia Karlsson & Vivan Franzen (2005). *The Phonology of Mongolian*. The Phonology of the World's Languages, Oxford University Press, Oxford.