**Intradimensional Bias in a Variable-free Model of Phonotactics**

*Brandon Prickett – bprickett@umass.edu – http://people.umass.edu/bprickett*

## 1. Introduction

a. Halle (1962) first proposed that assimilation and dissimilation patterns (e.g. vowel harmony, OCP effects, etc.) could be described using explicit, algebraic variables.
   i. This provided these kinds of patterns with simpler representations, which was useful since they're typologically common.
   ii. For example, if a language has voicing assimilation, this could be represented with the rule [-syllabic] → [αVoice] /_[αVoice], where [α] stands for either [+] or [-].
   iii. Variable can also be used with constraint-based representations, such as the markedness constraint *[αVoice][-αVoice] which would also enforce a voicing assimilation pattern.
   iv. So *[dt] and *[td] sequences would both violate the above constraint, since neither has matching voicing values for their first and second segment.

b. Hayes and Wilson (2008) proposed a model for phonotactic learning that did not include any explicit variables,[1] which inspired a range of research on whether they were necessary to include in models of phonology.[2]
   i. This model would need two constraints to represent the assimilation process above: *[+Voice][-Voice] and *[-Voice][+Voice].

c. A few different phenomena were used as evidence for algebraic representations (we'll just be focusing on Intradimensional Bias, but feel free to ask me about the others!):
   i. Identity Bias (Gallagher, 2013)
   ii. Identity Generalization (Berent, 2013; Berent, Wilson, Marcus, & Bemis, 2012; Gallagher, 2013)
   iii. Intradimensional Bias (Moreton, 2012)

d. Here, I'll show that you don't need variables to predict Intradimensional Bias.
   i. Instead, I use a novel mechanism, *Probabilistic Feature Attention*, paired with a variable-free phonotactic learner to model this phenomenon.
   ii. And I'll show that the mechanism can predict a behavior that variables don't: Similarity-based Generalization.

## 2. A Baseline Phonotactic Learning Model

a. The Hayes and Wilson (2008) phonotactic model learns a probability distribution over all possible words in a language after being trained on that language's lexicon.
   i. It represents this probability distribution using a set of weighted constraints like "*[+voice]" or "*[-tense][+word_boundary]".
   ii. The model's probability estimate for a word is proportional to the weighted sum of that word's constraint violations.[3]
   iii. They showed that the model's probability estimates for nonce words correlated well with speaker grammaticality judgements.

---

[1] Some earlier variable-free models of phonology did exist, such as Hare (1990) and Gasser and Lee (1992).

[2] The question as to whether variables should be included in models of cognition isn't limited to Phonology—see Marcus (2001) for a review of how these topics have been explored by psychologists.

[3] Specifically, $p(word_i) = \frac{e^{H_i}}{\sum e^{H_j}}$ where $H_i = \sum_{c \in C} w_c v_{c,i}$

b. In these simulations, I'll be using a different, but similar phonotactic model: GMECCS ("Gradual Maximum Entropy with a Conjunctive Constraint Schema"; Pater & Moreton, 2014; Moreton, Pater, & Pertsova, 2017).
c. The main difference between the two models is their constraint set.[4]
    i. Hayes and Wilson's learner induced a set of constraints that would allow it to describe the phonotactic patterns in a language…
    ii. …While GMECCS starts off with a constraint set that includes every possible conjunction of the relevant phonological features.
d. So if the only relevant features in a particular language were [voice] and [continuant] (e.g. if the only relevant segments were [d], [z], [t], and [s]) and words were only a single segment long, GMECCS would have a total of eight constraints:
    i. *[+voice], *[-voice], *[+continuant], *[-continuant]
    ii. *[+voice, +continuant], *[+voice, -continuant], *[-voice, +continuant], *[-voice, -continuant].
e. …And if that language only included the segments [d], [t], and [s], GMECCS might use the following weights to represent the language's lack of [z]:

| | ***[+vc.]** | ***[-vc.]** | ***[+cont.]** | ***[-cont.]** | ***[+vc., +cont.]** | ***[+vc., -cont.]** | ***[-vc., +cont.]** | ***[-vc., -cont.]** | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | 1 | 0 | 1.5 | 0 | 0 | 1 | *p* |
| **d** | * | | | * | | * | | | .32 |
| **z** | * | | * | | * | | | | .03 |
| **t** | | * | | * | | | | * | .32 |
| **s** | | * | * | | | | * | | .32 |

f. The model's learning process consists of finding weights like this for the constraints, so that the probability distribution predicted by the model matches the distribution of words in the language it's learning.
    i. For the results I present in §4-6, I used stochastic gradient descent, with batch sizes of 1 word (i.e. online gradient descent) to find optimal constraint weights.
    ii. However, the results are identical with standard (i.e. batch) gradient descent, which was used in past work with GMECCS (Moreton et al., 2017).
g. Crucially, GMECCS does not include any explicit algebraic variables—so just like the Hayes and Wilson (2008) model, it would require at least two constraints to describe an assimilation pattern: *[+Voice][-Voice] and *[-Voice][+Voice].

**3. Probabilistic Feature Attention (PFA)**

a. PFA has three main assumptions:
    i. When acquiring phonotactics, a learner doesn't attend to every phonological feature in every word.
    ii. This lack of attention creates ambiguity in the learner's input.
    iii. In the face of ambiguity, the learning algorithm errs on the side of assigning constraint violations.
b. This was inspired by Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014)—a mechanism used in NLP for regularizing neural networks.

[4] Another difference is that GMECCS uses both negative and positive weights for its constraints.

c. The idea that attention isn't equally distributed across features isn't new in cognitive science—for example, see Nosofsky's (1986) *Selective Attention*. The differences between this and PFA are:
   i. Attention is binary (features are either entirely attended to or entirely ignored)
   ii. The distribution of this attention is random (and resampled for each new word the learner is exposed to).

d. For example, let's consider the same simplified scenario where we only have four possible segments and two phonological features. When all features are attended to, we get unique violation profiles for all of the possible segments:

| | ***[+vc.]** | ***[-vc.]** | ***[+cont.]** | ***[-cont.]** | ***[+vc., +cont.]** | ***[+vc., -cont.]** | ***[-vc., +cont.]** | ***[-vc., -cont.]** |
|---|---|---|---|---|---|---|---|---|
| **d** | * | | | * | | * | | |
| **z** | * | | * | | * | | | |
| **t** | | * | | * | | | | * |
| **s** | | * | * | | | | * | |

e. But what if some features aren't attended to? Then the four segments will start to become ambiguous with one another.

f. Since PFA assigns violations in the face of ambiguity, an ambiguous segment will violate all of the constraints that the phonemes it's ambiguous between violate.

| **Attended Feature(s)** | | ***[+vc]** | ***[-vc]** | ***[+cont]** | ***[-cont]** | ***[+vc, +cont]** | ***[+vc, -cont]** | ***[-vc, +cont]** | ***[-vc, -cont]** |
|---|---|---|---|---|---|---|---|---|---|
| **[voice]** | **T** | | * | * | * | | | * | * |
| | **D** | * | | * | * | * | * | | |
| **[cont.]** | **ð** | * | * | | * | | | * | * |
| | **Z** | * | * | * | | * | * | | |
| **None** | **?** | * | * | * | * | * | * | * | * |

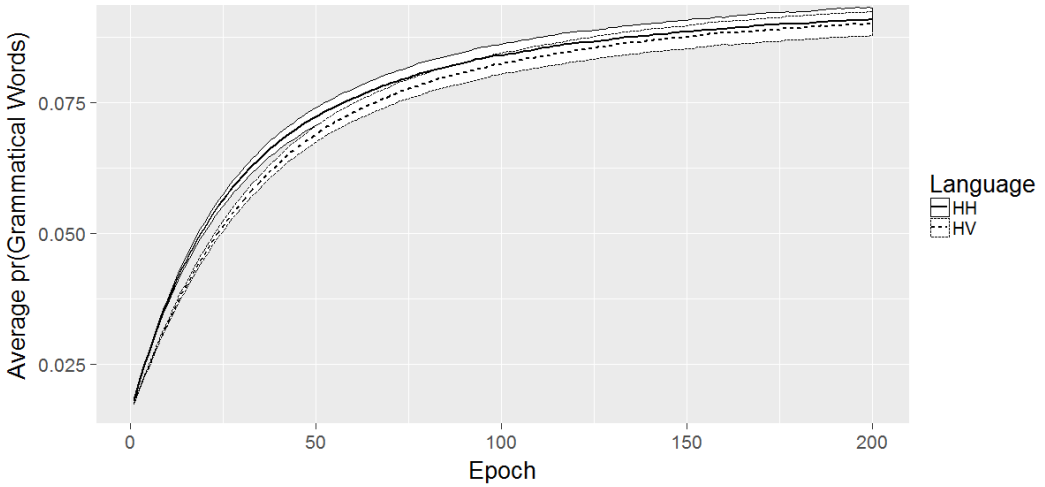
g. Over the course of acquisition, each new piece of learning data (i.e. each new word) is probably only going to give the learner a partial picture of the overall pattern (since none, some, or all of the features may be ignored for that datapoint).

**4. Intradimensional Bias**

a. Moreton (2008, 2012) showed that assimilation patterns like "agree in height" were easier to learn than more arbitrary patterns like "voiced consonants must occur before high vowels" (that is, an *Intradimensional Bias*).

b. Moreton (2012) showed that this bias isn't predicted in a phonotactic learner without variables (although, see Doucette, 2017)—since a pattern like height agreement (HH) takes the same number of constraints to represent as the height-voicing pattern (HV), they're learned at the same rate.
   i. HH = *[-high][+high], *[+high][-high]
   ii. HV = *[+high][-voice], *[-high][+voice]

c. I replicated this result with GMECCS:



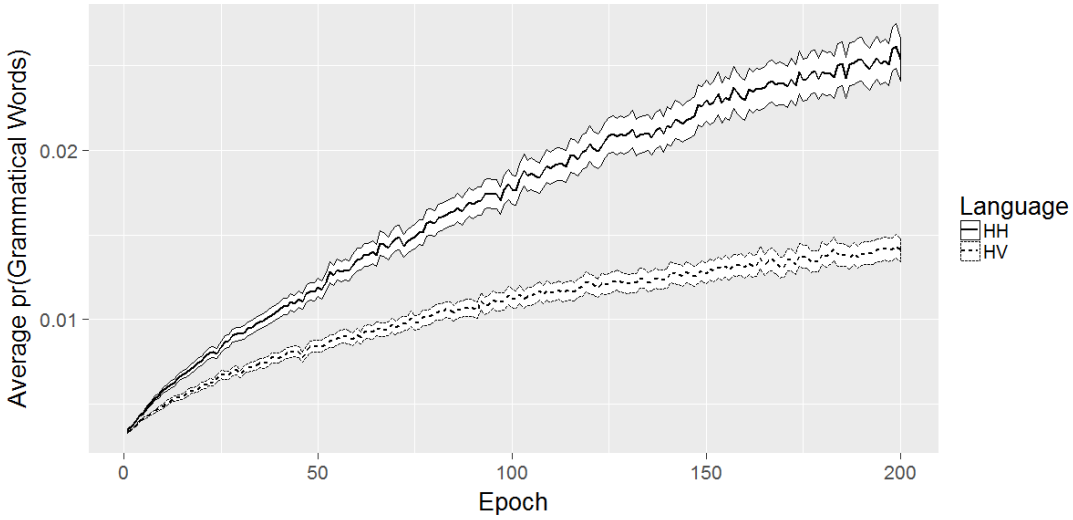

d. While the HH pattern shows a slight advantage early on for GMECCS, there's a large amount of overlap between the two curves' standard error for most of learning.

e. However, when Moreton (2012) added a constraint with variables (*[αHigh][αHigh]), his model was able to predict Intradimensional Bias. He took this as evidence that variables are necessary in Phonology.

**5. PFA and Intradimensional Bias**

a. Can PFA predict this bias without variables?

b. To test this, I replicated Moreton's (2012) simulation, but with a version of GMECCS equipped with PFA (probability of each feature being attended to on each update = .25):




c. This shows that a model *can* capture Intradimensional Bias without variables, as long as it has PFA.

   i. PFA is able to capture the phenomenon because the more features that are relevant to a pattern, the more likely the pattern is to be obscured by random ambiguity.

   ii. This means that HH (which only involves one feature) will be obscured less often than HV (which involves two features).

**6. Similarity-based Generalization**

a. So far I've focused on showing that PFA can predict the same kind of behavior as variables, but are there any predictions that PFA and variables differ in?

b. To see a way in which the predictions differ between the two proposals, we'll look at a phenomenon that I'll call *Similarity-based Generalization*.

i. Standard theories of phonological generalization typically predict that patterns are generalized to novel sounds within a natural class (Halle, 1978).

ii. For example, if speakers observe a pattern involving a set of voiced sounds, they should only generalize that pattern to other voiced sounds.

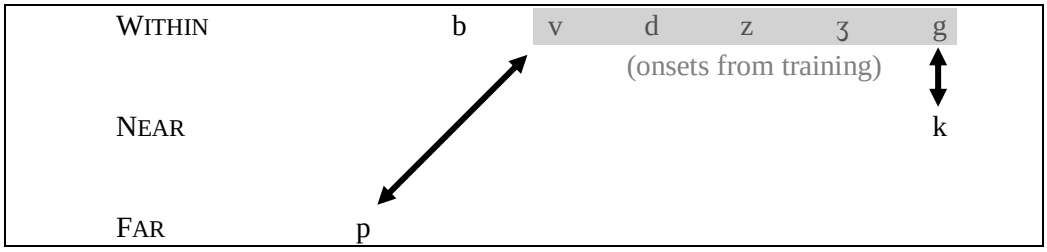
iii. However, another possibility is that generalization is similarity-based. If this is the case, then language learners generalize patterns to any sound that's similar to the set of sounds that undergo the pattern (Mielke, 2008).

c. Cristia et al. (2013) observed Similarity-Based Generalization in an experiment in which participants learned a phonotactic pattern that limited words' onsets to a particular natural class.

i. For example, participants might be trained on words with onsets that were all [+voice].

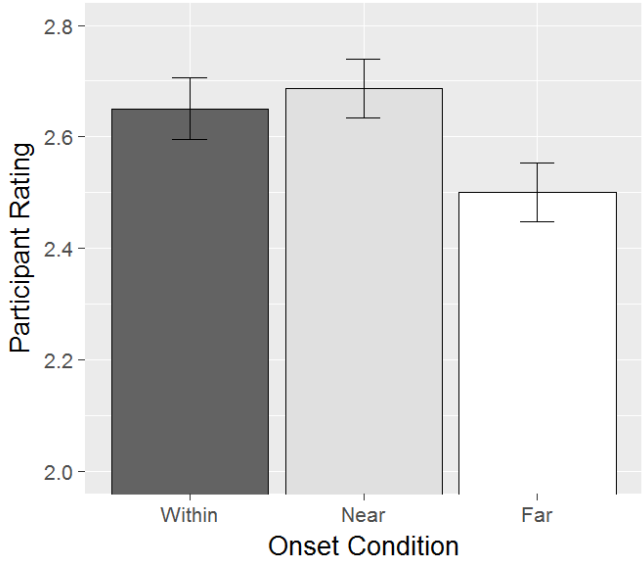
ii. They then tested participants on three kinds of data that weren't present in training:

- WITHIN: these were words whose onset was withheld from training but was within the natural class of legal onsets. For example, if subjects were trained with voiced onsets, the WITHIN test item could be [b].
- NEAR: these were words whose onset was close (both featurally and phonetically) to the class of legal onsets in training, but that weren't within the relevant natural class. For example, if participants were trained on the voiced segments [v], [d], [z], [g], and [ʒ], a segment that would be relatively close to that group would be [k] (since it's only 1 feature away from [g] and only two away from most others).
- FAR: these are words whose onsets are not within or near the class of onsets from training. In the example above, [p] would be considered FAR because there are no labial stops present in the training data.

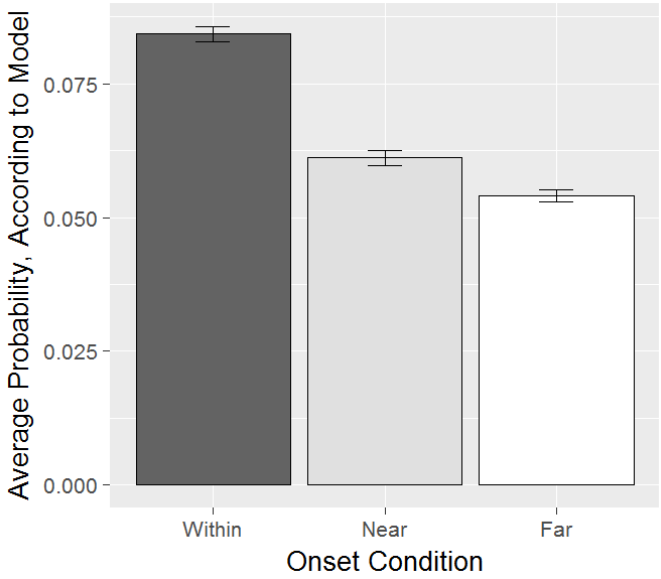



ii. If generalization was based on natural classes, participants should only rate WITHIN words as being grammatical, but if generalization is similarity-based, then both WITHIN and NEAR words should be considered grammatical.

d. Cristia et al.'s (2013) experiment results showed that their participants assigned more probability to WITHIN and NEAR words than to their FAR counterparts, but there was no significant different between the former two groups:

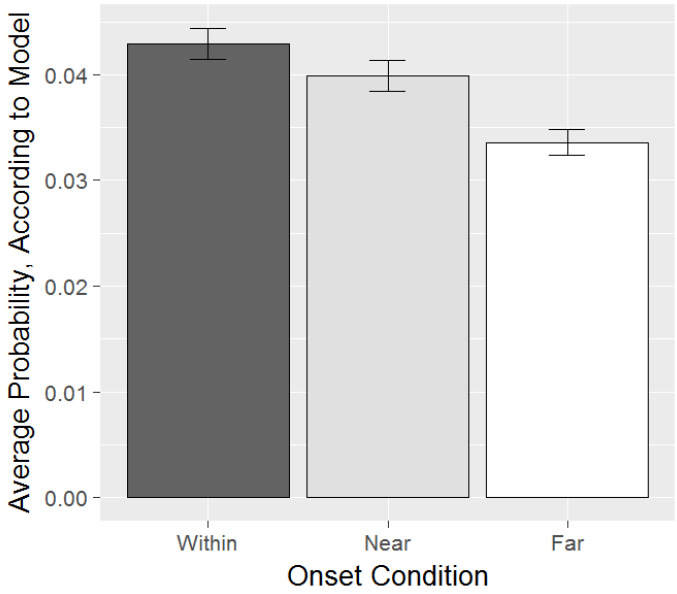



e. If the standard version of GMECCS is given the same kind of training data as their participants, it is unable to give as much probability to the NEAR category of stimuli as it does to the WITHIN category (i.e. it generalizes in a relatively natural-class based way, since its constraints are based on natural classes):




f. Variables won't change the prediction that GMECCS makes, since this pattern is within a single segment (the words' onsets) and variables need to be restricted to occurring on the same feature across segments.
    i. If not, you could represent relatively strange patterns like "in onsets, make the value for [voice] match the value for [continuant]" with a single constraint like "*[αVoice, αContinuant]".

g. However, when GMECCS is equipped with PFA, Similarity-based Generalization is correctly predicted—with a similar amount of probability being given to both the WITHIN and the NEAR categories:




h. This is because over the course of learning, NEAR words are likely to become ambiguous with words from training (since they share a relatively large number of features), which causes the model to assign both more probability to the NEAR category than to FAR.

**7. Conclusions**

a. Here, I showed that Intradimensional Bias can be accounted for without variables—all you have to do is assume that ambiguity happens in a random-but-structured way throughout the learning process (i.e. that a learner uses PFA).

b. I also showed that unlike variables, PFA provides a unified account of Intradimensional Bias and Similarity-based Generalization.

c. There's a lot more to look at in regards to PFA:

   i. Do child acquisition errors look like the kinds of errors PFA predicts?

   ii. How does PFA do when it's scaled up to real-language problems? For example, Berent et al. (2012) used variables to model an identity pattern in Hebrew.

   iii. Are there other ways in which the predictions made by PFA and variables differ?

d. But for now, it seems like PFA could be a better way to handle the phenomena that have been used as evidence for variables in Phonology.

**References**

Berent, I. (2013). The phonological mind. *Trends in Cognitive Sciences*, *17*(7), 319–327.

Berent, I., Wilson, C., Marcus, G., & Bemis, D. K. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry*, *43*(1), 97–119.

Cristia, A., Mielke, J., Daland, R., & Peperkamp, S. (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, *4*(2), 259–285.

Doucette, A. (2017). Inherent Biases of Recurrent Neural Networks for Phonological Assimilation and Dissimilation. *ArXiv Preprint ArXiv:1702.07324*.

Gallagher, G. (2013). Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, *30*(2), 253–295.

Gasser, M., & Lee, C.-D. (1992). Networks that learn about phonological feature persistence. In *Connectionist Natural Language Processing* (pp. 349–362). Springer.

Halle, M. (1962). A descriptive convention for treating assimilation and dissimilation. *Quarterly Progress Report*, *66*, 295–296.

Halle, M. (1978). *Knowledge unlearned and untaught: What speakers know about the sounds of their language*.

Hare, M. (1990). The role of trigger-target similarity in the vowel harmony process. *Annual Meeting of the Berkeley Linguistics Society*, *16*, 140–152.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Marcus, G. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford University Press.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*(1), 83–127.

Moreton, E. (2012). Inter-and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*(1), 165–183.

Moreton, E., Pater, J., & Pertsova, K. (2017). *Phonological Concept Learning*.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Pater, J., & Moreton, E. (2014). Structurally biased phonology: complexity in learning and typology. *The EFL Journal*, *3*(2).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.