# Similarity-based Phonological Generalization[*]

**Brandon Prickett**
University of Massachusetts Amherst
bprickett@umass.edu

## 1 Introduction

Models of phonological learning typically motivate the generalization of patterns using abstract representations that can refer to an entire class of sounds. Here, I present a computational model of an alternative approach to generalization based on featural similarity that more accurately predicts the results of an experiment by Cristia et al. (2013).

Halle (1978) showed assimilation of the possessive suffix in English is generalized by speakers to non-English segments, such as [x] (this is often referred to as the *Bach Test*, since the [x]-final word Halle used was *Bach*). He suggested that an abstract, partial featural description, as exemplified in (1), could explain this.

(1) All [-voice] segments trigger
assimilation to [-voice].

The use of the feature bundle [-voice] to refer to all sounds that are voiceless gives this representation the ability to generalize to the segment [x], regardless of whether speakers have ever been exposed to it (since [x] is voiceless). Henceforth, I will call this *abstraction-based generalization*, since a novel segment is included in a pattern due to that pattern's abstract representation.

An alternative explanation for phonological generalization is that it's the result of "cognitive biases emergent from online calculations of similarity"[1] (Cristia et al. 2013:279). In the context of assimilation from above, this might look like the following:

(2) Likely to trigger
↑ [p, f, θ, k, …] (*attested triggers)*
[ɸ, p̪, t̪, x, …] *(unattested, similar)*
↓ [b, v̆, ð, g, …] *(dissimilar segments)*
Unlikely to trigger

Henceforth, I will refer to the paradigm in (2) as *similarity-based generalization*. In similarity-based generalization, the representation of a pattern is not what causes generalization to novel sounds. Instead, the novel sounds' similarity to attested sounds biases speakers toward treating them in a similar way, which causes patterns to generalize to novel, similar segments.

To test which theory better predicts how humans generalize phonological patterns, I created a Maximum Entropy (MaxEnt) phonological learner that uses similarity in its learning update to encourage the generalization of patterns to featurally similar segments. My learner's predictions match the results of Halle's (1978) Bach Test, as well as more recent experimental results from Cristia et al. (2013). My similarity-based learner predicts the human behavior better than a minimally different, previously proposed learner that relies on abstraction-based generalization.

[*] Thanks to Joe Pater, Andrew Lamont, the UMass Phonology Reading Group, and the SCiL reviewers for helpful discussion and suggestions. Also, thanks to Gaja Jarosz and John Kingston for advising a related project that led to this paper. This research was supported by NSF Grant #BCS-1650957.

[1] While Cristia et al. (2013) suggested that articulatory similarity might be a better predictor for generalization than featural similarity, here I focus on the latter and leave exploring the former to future work.

## 2 Modeling generalization with GMECCS

I used GMECCS[2] (Moreton et al. 2017) to model abstraction-based generalization. GMECCS uses a gradient descent learning algorithm (with a single parameter $\eta$, representing the model's learning rate) and a constraint set that includes all possible conjunctions of the features of interest. For example, if [voice] and [continuant] are the only contrastive features necessary in a simulation (e.g. if the only relevant sounds are [z], [d], [s], and [t]), and the words have a maximum length of 1 segment, the constraint set for GMECCS would be: *[+voice], *[-voice], *[+cont.], *[-cont.], *[+voice,+cont.], *[+voice,-cont.], *[-voice,+cont.], and *[-voice,-cont.]. Moreton et al. (2017) used GMECCS to properly predict the relative learnability of different phonotactic patterns.[3] However, GMECCS is also useful for testing abstraction-based generalization, since a subset of its constraints use abstract featural descriptions to refer to multiple sounds (e.g. *[+voice] which applied to both [z] and [d] in the example above, and could generalize to any voiced segment).

## 3 Modeling generalization with Sim-Gen

In order to compare the two types of generalization discussed in §1, I created a MaxEnt learning model that uses similarity-based generalization (henceforth, this model will be called *Sim-Gen*). Sim-Gen differs from GMECCS in only two ways: its constraint set and its update algorithm. Sim-Gen's constraints do not represent every possible combination of features. Instead, they only represent every possible feature bundle that refers to a single segment. So for the example in §2, the constraint set would be: *[+voice,+cont.], *[+voice,-cont.], *[voice,+cont.], and *[-voice,-cont.]. Because these constraints don't abstract away from individual segments (e.g. through the use of partial featural descriptions), they won't lead to abstraction-based generalization.

Instead, similarity-based generalization results from the learner's update algorithm. At each epoch, the constraint weights are updated to better reflect the training data (see Morteon et al. 2017 for more on this step). However, this learner differs from GMECCS in that each change to a constraint's weight also "leaks" onto all of the other constraints. These leaks are larger for constraints that are more similar to the original constraint, which makes the learner biased toward assigning similar weights to constraints that are similar to one another. This is formalized in Equations (1-2):

$$\delta \mathrm{w}_j = \theta \left[ \frac{\Delta \mathrm{w}_i}{\mathrm{dist}(c_i, c_j)} \right] \quad (1)$$

$$dist(c_i, c_j) = |Features\ c_i\ \&\ c_j\ differ\ in| \quad (2)$$

Where:
$c_i$ is the primary constraint being updated,
$c_j$ is the constraint being leaked on,
$\Delta w_i$ is the primary update (to $c_i$'s weight),
$\delta w_j$ is the leaked update (to $c_j$'s weight),
And $\theta$ is a parameter controlling leak size.

In the equations above, every constraint $c_j$ that isn't undergoing the primary update (i.e. the update $\Delta w_i$ that's based on the learning data) undergoes a leaked update $\delta w_j$ that's proportional to $\Delta w_i$ and inversely proportional to that constraint's feature distance from the primary constraint $c_i$ (where distance is the number of features the constraints differ in). This results in constraints having high weights not only when they help to describe the learning data (due to $\Delta w_i$), but also when they happen to be similar to the constraints that help describe the learning data (due to $\delta w_j$). This, coupled with constraints representing all relevant segments (including those that might be unattested in a language, such as [x] in English), results in similarity-based generalization.[4]

## 4 Modeling the Bach Test

To test whether both models predicted the kind of generalization observed by Halle (1978), I trained them on a toy language that was made to represent the parts of English relevant to the Bach Test. The toy language's segment inventory consisted of the set: [d], [t], [g], [k], [z], and [s]. In addition to these six segments, both models had constraints referring to the velar fricatives [ɣ] and [x]. Since GMECCS

[2] GMECCS is an acronym for "Gradual Maximum Entropy with a Conjunctive Constraint Schema."

[3] My model (described in §3) also predicts these relative learnabilities.

[4] See Rumelhart and McClelland's (1987) "blurring" for an alternative approach to generalization that isn't abstraction-based.

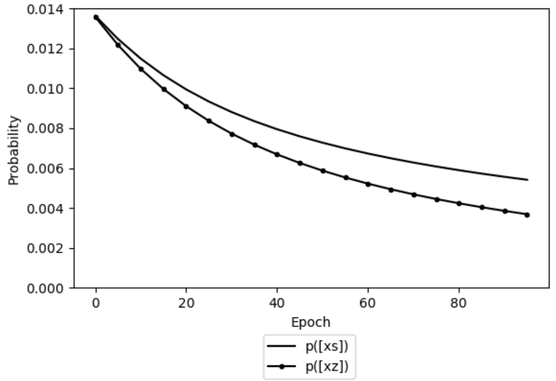

**Figure 1:** Bach Test simulation with the abstraction-based GMECCS learner: 100 epochs with η=.01

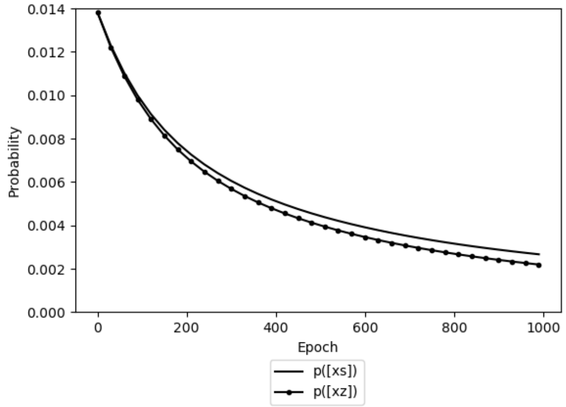

**Figure 2:** Bach Test simulation with the similarity-based Sim-Gen learner: 1000 epochs with η=0.01, θ=0.5

and Sim-Gen only model phonotactics, the simulations described here modeled the learning of a phonotactic restriction against bigrams[5] that disagreed in voicing (e.g. *[kz]), rather than voicing assimilation in the context of a morpho-phonological alternation. The probabilities that the models gave to segment types over the course of learning are shown in Figures 1 and 2. Both models acquired a preference for clusters that agreed in voicing,[6] even when a member of those clusters was unattested in the training data. That is, they both learned that p([xs]) > p([xz]), despite [x] being unattested. This demonstrates that both Sim-Gen and GMECCS can successfully simulate a kind of generalization that's analogous to Halle's (1978) Bach Test.

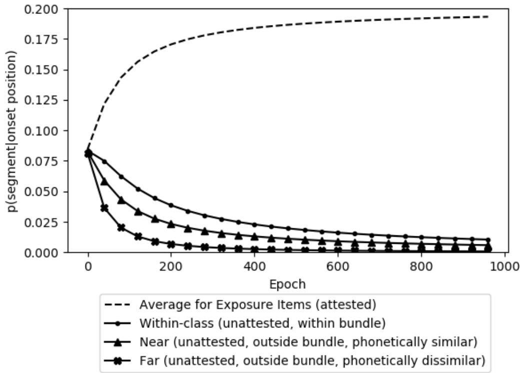

**Figure 3:** Experiment simulation, with the abstraction-based GMECCS learner: 1000 epochs, η=.01

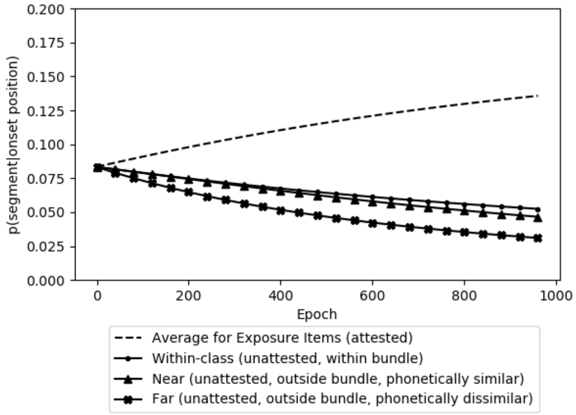

**Figure 4:** Experiment simulation, with the similarity-based Sim-Gen learner: 1000 epochs, η=0.01, θ=0.5

## 5 Modeling Cristia et al.'s (2013) data

In a study by Cristia et al. (2013), participants generalized an onset restriction from an artificial language to novel words with attested onsets (their EXPOSURE condition). To a lesser degree, subjects also generalized this restriction to segments that weren't in training but that were within the EXPOSURE segments' feature bundle (their WITHIN condition), and to segments that were phonetically similar to EXPOSURE items (their NEAR condition). However, subjects didn't generalize to segments that were phonetically dissimilar to EXPOSURE segments (their FAR condition). I ran simulations of their experiment using both GMECCS and Sim-Gen. Figures 3 and 4 illustrate the probability that

[5] This required the models to have unigram and bigram constraints. In order to find the value of dist($c_{\text{unigram}}$, $c_{\text{bigram}}$), Sim-Gen compares the unigram constraint's segment to the segment in the bigram constraint that it's most similar to. The final result is the distance between these segments, plus 0.5 (to penalize the difference in constraint lengths).

[6] Sim-Gen took longer to acquire this preference. However, since the crucial comparison for these simulations is whether Sim-Gen treats p([xs]) and p([xz]) differently at any point in its learning curve, speed of acquisition isn't relevant.

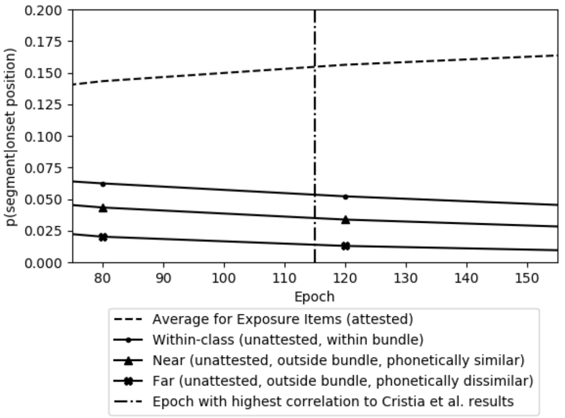

**Figure 5:** Point in learning where the abstraction-based GMECCS learner correlates the most with Cristia et al.'s (2013) experimental data. Full simulation in Figure 3.

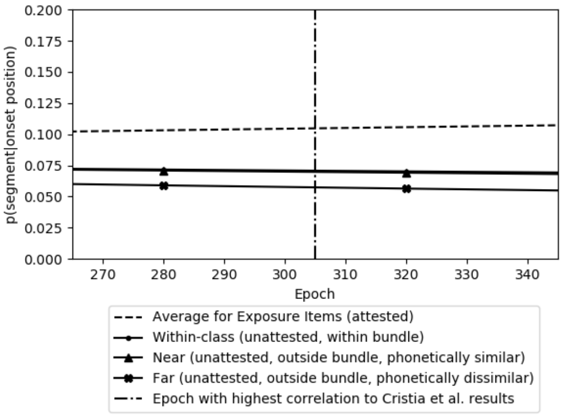

**Figure 6:** Point in learning where the similarity-based Sim-Gen learner correlates the most with Cristia et al.'s (2013) experimental data. Full simulation in Figure 4.

the learners assigned to each segment type appearing in onset position over the course of learning. GMECCS consistently treated the WITHIN and NEAR categories differently, while the similarity-based learner treated them almost identically for a significant portion of learning. This shows that Sim-Gen models the human behavior more accurately than GMECCS.

Since Cristia et al.'s results represent a single point in their subjects' learning, another way of examining these simulations is to find which point in each model's learning curve best fits the human behavior and comparing those points across models. For GMECCS, the model's assigned probabilities had the highest correlation with subjects' responses on the 115th epoch (Pearson's *r*=.986). For Sim-Gen, the most correlated epoch was the 305th (Pearson's *r*=.999). Figures 5 and 6 show the segment probabilities for each learner at its most correlated point.

Even when it best matches the human data, GMECCS assigns different probabilities to NEAR and WITHIN segments, while Sim-Gen assigns them almost equal probabilities. In addition, the probability that Sim-Gen gives FAR segments is visibly lower at this point than the other two categories, which matches the Cristia et al. (2013) data well. GMECCS, on the other hand, only assigns similar probabilities to WITHIN and NEAR segments at the beginning of learning (not shown in the figures), when all segments (including FAR) are assigned a similar probability.

## 6 Conclusions

Halle (1978) and Cristia et al. (2013) both observed generalization of phonological patterns. In Halle's (1978) Bach Test, a phonological alternation was triggered both by attested voiceless segments and the unattested, voiceless sound [x]. In Cristia et al.'s (2013) experimental results, subjects generalized an onset restriction both to unattested segments sharing a feature with the attested onsets and to unattested segments that were similar to the attested onsets. The simulations ran in this study showed that both abstraction-based and a similarity-based models predict the generalization described by Halle (1978). However, only Sim-Gen (the similarity-based learner) predicts the kind of generalization that was observed by Cristia et al. (2013). This supports the idea that human generalization is grounded in similarity, rather than abstract, partial feature representations.

## References

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

Halle, M. (1978). Knowledge unlearned and untaught: what speakers know about the sounds of their language. In: M. Halle, J. Bresnan & G.A. Miller (eds.), *Linguistic theory and psychological reality*. MIT Press, Cambridge, MA, 294-303.

Moreton, E., J. Pater, & K. Pertsova (2017). Phonological concept learning. *Cognitive science, 41(1)*, 4-69.

Rumelhart, D. E., & J. L. McClelland (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing. *Mechanisms of language acquisition*. Lawrence Erlbaum Associates, Hillsdale, NJ, 195-248.