Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias

To appear in *Language*

James White

Department of Linguistics

University College London

Email: j.c.white@ucl.ac.uk

Mailing address:

Chandler House
2 Wakefield Street
London WC1N 1PF
United Kingdom

ABSTRACT

Saltatory alternations occur when two sounds alternate with each other, excluding a third sound that is phonetically intermediate between the two alternating sounds (e.g. [p] alternates with [β], with non-alternating, phonetically intermediate [b]). Such alternations are attested in natural language, so they must be learnable; however, experimental work suggests that they are dispreferred by language learners. This paper presents a computationally implemented phonological framework that can account for both the existence and the dispreferred status of saltatory alternations. The framework is implemented in a maximum entropy learning model (Goldwater & Johnson 2003) with two significant components. The first is a set of constraints penalizing correspondence between specific segments, formalized as *MAP constraints (Zuraw 2007, 2013), which enables the model to learn saltatory alternations at all. The second is a substantive bias based on the P-map (Steriade 2009 [2001]), implemented via the model's prior probability distribution, which favors alternations between perceptually similar sounds. Comparing the model's predictions to results from artificial language experiments, the substantively biased model outperforms control models that do not have a substantive bias, providing support for the role of substantive bias in phonological learning.*

Keywords: phonology, substantive bias, learnability, maximum entropy, modeling, P-map

* For helpful comments and discussion, I would like to thank Bruce Hayes, Megha Sundara, Kie Zuraw, Robert Daland, Sharon Peperkamp, Adam Albright, Jennifer Culbertson, two anonymous reviewers, the associate editor Megan Crowhurst, members of the UCLA phonology seminar, and audiences at the 38th Boston University Conference on Language Development and the 88th Annual Meeting of the LSA.

1. INTRODUCTION. Ultimately, the goal of phonological theory is not just to account for which patterns are possible or impossible in human language; the theory must also be able to explain why certain patterns, though possible, are more difficult to learn (or are otherwise dispreferred) compared to others. This idea has a long history (e.g. see Kiparsky 1982:59–60), but until recently most of the theoretical work in phonology focused on the first issue: developing a framework that could derive the existing patterns while excluding the ones deemed impossible. In recent years, however, phonology has experienced an explosion of new empirical results thanks, in part, to new methods for collecting data and more sophisticated ways of implementing theories in computational models. These advances have provided us with new ways of exploring the learnability of phonological patterns.

An area of research that has greatly benefitted from these advances, and one that forms the basis of the current study, is investigating which *a priori* biases the learner brings to the language acquisition process. Research on this topic, traditionally expressed in terms of discovering the nature of Universal Grammar, has long focused on language analysis. With careful analysis of individual languages, followed by a comparison of such analyses across many languages, phonological theories can be tested on their ability to predict the observed typology: existing languages should be derivable using the theoretical framework whereas unattested languages (at least, ones deemed to be impossible) should not. A clear application of this approach can be seen in calculating a factorial typology in Optimality Theory (OT; e.g. see Prince & Smolensky 2004 [1993], Kaun 1995, Gordon 2002, Baković 2004, Zuraw 2010).

Recent experimental and computational work has provided new insights on the issue of biases in phonological learning by focusing more directly on the learner rather than solely on the analysis of the final, adult grammar. Results in this area have come from a variety of studies, including artificial language experiments (Pycha et al. 2003, Peperkamp et al. 2006b, Wilson 2006, Peperkamp & Dupoux 2007, Moreton 2008, Carpenter 2010, Nevins 2010, Skoruppa & Peperkamp 2011, Skoruppa et al. 2011, Baer-Henney & van de Vijver 2012, Finley & Badecker 2012, White 2014), studies comparing corpus analysis with native speaker intuitions (Zuraw 2007, Hayes et al. 2009, Becker et al. 2011, Becker et al. 2012, Hayes & White 2013), infant experiments (Seidl & Buckley 2005, Cristià & Seidl 2008, Chambers et al. 2011, White & Sundara 2014), and computational modeling (Peperkamp et al. 2006a, Wilson 2006, Calamaro & Jarosz 2015).

This growing body of work has been fruitful, but many questions remain. A particular area of interest is determining whether phonetic substance plays a role during phonological learning, and if so, understanding how to characterize its role. There is little doubt that the phonetic properties of speech sounds play an important role in shaping the phonologies of the world's languages. Under some accounts, learners have access to these phonetic properties as part of their subconscious linguistic knowledge, and this knowledge plays a direct role in biasing their learning; phonological patterns with strong phonetic motivation are favored by learners relative to patterns without such motivation (e.g. Archangeli & Pulleyblank 1994, Hayes 1999, Hayes & Steriade 2004, Steriade 2009 [2001]). Inductive biases that specifically draw on prior phonetic knowledge have been called SUBSTANTIVE BIASES (Wilson 2006). An opposing view holds that phonetic substance has little to no role to play in synchronic phonological learning, and that the role of phonetics in shaping phonological systems is limited almost entirely to the domains of production and perception (i.e. errors of transmission; see Ohala 1981, 1993; Hale & Reiss 2000; Blevins 2004; Yu 2004; see also Moreton 2008). On the experimental side, several researchers have appealed to a substantive bias account to explain their results (e.g. Wilson 2006, Finley 2008, Baer-Henney & van de Vijver 2012, White 2014). However, other studies looking for effects of substantive bias in experiments have found null results, and the overall evidence on the issue of substantive biases in phonological learning remains inconclusive (for a review, see Moreton & Pater 2012).

In this paper, my starting point is a phonological phenomenon—saltation—that is attested in natural languages (Hayes & White 2015), but is dispreferred by human learners in experiments (White 2014, White & Sundara 2014). Phenomena with these two properties pose a challenge for traditional phonological theory; we must have a framework that can simultaneously account for the fact that a pattern is learnable (because it is attested) and the fact that it is dispreferred by learners.

I argue for a theoretical framework with the following notable properties. First, the grammar is augmented with a set of constraints, formalized as *MAP constraints (Zuraw 2007, 2013), which are capable of penalizing correspondences between any two individual sounds (as opposed to traditional faithfulness constraints that only penalize changes at the level of the feature); these constraints make it possible for saltation to be derived at all. Second, I adopt a substantive bias, based on the principles embodied in the theory of the P-map (Steriade 2009 [2001]), which causes the learner to favor alternations between perceptually similar sounds relative to alternations between dissimilar sounds. Finally, following an approach pioneered by Wilson (2006), the framework is implemented in a maximum entropy (MaxEnt) learning model (Goldwater & Johnson 2003), which allows the substantive bias to be implemented as a 'soft' bias via the model's prior probability distribution. To test the model, I compare its predictions to results from artificial language experiments. Overall, I show that a model with these components is successful at predicting the desired learning behavior: saltations are initially dispreferred, but with sufficient input data, they can eventually be learned. I further show that the substantively biased model provides a better fit to the experimental data than control models without a substantive bias, providing support for the idea that substantive bias plays a role during phonological learning.

The paper is organized as follows. The next section provides background on saltation and the challenge that it poses for phonological theory. §3 presents the proposed solution: an implemented MaxEnt model with a P-map bias. In §4, I provide an overview of experimental data from White 2014 showing that saltatory alternations are dispreferred by language learners; these data will serve as the basis for testing the model. §§5–7 focus on testing the model's predictions. The remaining sections (§§8-10) discuss outstanding issues and conclude.

2. THE PROBLEM OF SALTATORY ALTERNATIONS. Hayes and White (2015:267) define SALTATION in terms of features. An alternation, A ~ C, is considered saltatory if there exists (in the phonological inventory of the language) some segment, B, such that B is left unchanged in the context where A alternates with C; and 'for every feature for which A and C have the same value, B also has that value, but that B differs from both A and C'. Another way of expressing this definition is in terms of a subset relationship: if the set of features that differ between A and B, and the set of features that differ between B and C, are each subsets of the set of features that differ between A and C, then the alternation, A ~ C, is saltatory.

Figure 1 shows an example of a saltation found in the Campidanian dialect of Sardinian (see Bolognesi 1998). In this language, voiceless stops are realized as voiced fricatives in intervocalic contexts (compare: [**p**ãi] 'bread', [sːu **β**ãi] 'the bread'), but voiced stops are unchanged in intervocalic position ([**b**ĩu] 'wine', [sːu **b**ĩu] 'the wine', *[sːu **β**ĩu]). In this example, [p] and [b] differ only in [voice], [b] and [β] differ only in [continuant], and [p] and [β] differ in both [voice] and [continuant]; thus, the [p ~ β] alternation saltates over phonetically intermediate, non-alternating [b].

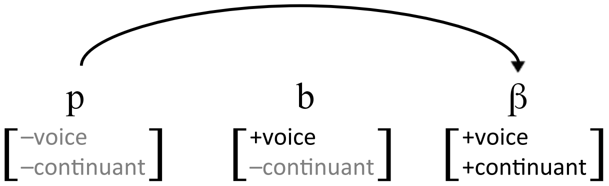

FIGURE 1. A saltatory alternation in Campidanian Sardinian.

It is worth noting that the definition of saltation offered by Hayes and White (2015) can be generalized to refer to any measure of similarity (as opposed to just features) as in (1).

(1) Definition of saltation from Hayes & White 2015, generalized.

a. Let A, B, and C be phonological segments. Assume some relevant measure of similarity, S.

b. Suppose that S(A, B) > S(A, C) and S(B, C) > S(A, C); that is, the similarity of A and B and the similarity of B and C are each greater than the similarity of A and C.

c. If in some context A alternates with C, but B is unchanged, then the alternation A ~ C is a saltation.

The measure of similarity referenced in (1) could be number of shared features, but it could also refer to perceptual similarity, articulatory similarity, and so on.

It is difficult to be sure about the cross-linguistic prevalence of saltatory alternations as there has not (to my knowledge) been a quantitative typological study. Hayes and White (2015) describe several cases of saltation, arguing that they only seem to arise through historical accident, and that they may be unstable when they do arise. As far as I am aware, the Campidanian Sardinian case just discussed is the only case of saltation that (a) is fully productive, (b) applies to a full natural class of sounds (as opposed to a single sound in isolation), and (c) is not limited to a specific morphological context.[1] Thus, the available evidence suggests that saltatory phonological patterns which are regular, productive, and stable are relatively rare (though not unattested).

Saltatory alternations, like the one presented in Figure 1, cannot be derived in ‘classical OT’, meaning the theory as originally proposed by Prince and Smolensky (2004 [1993]), augmented by Correspondence Theory (McCarthy & Prince 1995). The problem, which has been pointed out by others (Lubowicz 2002, Ito & Mester 2003, McCarthy 2003, Hayes & White 2015), stems from the excessive nature of the change involved in such alternations. In order for /p/ to surface as [β], the markedness constraints *V[−voice]V and *V[−continuant]V must each be ranked higher than the opposing faithfulness constraints, IDENT(voice) and IDENT(continuant), respectively, as in (2). However, this ranking would also result in /b/ surfacing as [β], as shown in (3). In order to protect /b/ from changing to [β], IDENT(continuant) would need to outrank *V[−continuant]V; however, this would also prevent /p/ from changing all the way to [β] – it would instead ‘get stuck’ at intermediate [b]. Thus, if /b/ surfaces faithfully as [b], but /p/ saltates over [b] to surface as [β], we get a ranking paradox in classical OT.

(2) Tableau showing /VpV/ ⟶ [VβV] in classical OT.

/VpV/	*V[−voice]V	*V[−cont]V	IDENT(cont)	IDENT(voice)
☞ VβV			*	*
VbV		*!		*
VpV	*!	*		

(3) The same ranking incorrectly derives /VbV/ ⟶ *[VβV].

/VbV/	*V[−voice]V	*V[−cont]V	IDENT(cont)	IDENT(voice)
☞ *VβV			*	
☹ VbV		*!		

[1] For instance, a reviewer asks whether there is evidence in English for saltatory alternations. English does have the process of velar softening, which results in [k ~ s] alternations (e.g. *opaque* [oʊˈpeɪk], *opacity* [oʊˈpæsɪɾi]) that saltate over [t]. Note, however, that the process is lexically and morphologically limited, and that it has somewhat limited productivity (Pierrehumbert 2006).

2.1. MODIFYING THE CONSTRAINT SET: *MAP CONSTRAINTS. The real problem underlying classical OT's inability to generate saltation is the set of traditional feature-based faithfulness constraints assumed in the theory, such as IDENT(cont) and IDENT(voice). Without modifying the constraint set, the problem persists even if we abandon strict dominance in favor of weighted constraints (as we will do below when we switch to maximum entropy grammars; see §3). With traditional IDENT constraints, it is necessarily the case that that a longer journey (e.g. /p/ ⟶ [β]) will be less harmonic than a shorter journey (e.g. /b/ ⟶ [β]), assuming that the two outputs are identical and therefore violate the same markedness constraints.

For saltation to be allowed, the opposite must be possible; that is, it must be possible for a short journey (e.g. /b/ ⟶ [β]) to incur a greater penalty than a long journey (e.g. /p/ ⟶ [β]). As a solution, following Hayes and White (2015), I adopt the *MAP family of correspondence constraints, proposed by Zuraw (2007, 2013). Unlike traditional IDENT constraints, *MAP constraints are not restricted to penalizing changes of a single feature. Instead, they penalize correspondences between any two natural classes of sounds. The constraints are formalized as in (4), adapted from Zuraw 2007, 2013.

(4) *MAP, formalized[2]

*MAP(*x*, *y*): violated when a sound that is a member of natural class *x* corresponds to a sound that is a member of natural class *y*.

For the cases considered here, what will be necessary are segment-specific versions of the constraints. For instance, *MAP(p, β) would be violated when [p] is in correspondence with [β]. In this case, the constraint *MAP(p, β) may be considered notational shorthand for $\text{*M}{\small\text{AP}}\left(\begin{bmatrix}-\text{voice}\\-\text{cont}\\+\text{labial}\end{bmatrix}, \begin{bmatrix}+\text{voice}\\+\text{cont}\\+\text{labial}\end{bmatrix}\right)$, where each of the natural classes happens to consist of only a single sound.

[2] Zuraw's formalism also allows the constraints to specify a particular context in which a pair of sounds must not be in correspondence (e.g. a sound of natural class *x* in context A__B should not correspond to a sound of natural class *y* in context C__D). Context-specific versions of the constraints are not necessary here.

Before moving on, I will clarify a couple of working assumptions about the *MAP constraints. First, I follow Zuraw (2013) in assuming that the *MAP constraints should be limited to evaluating correspondences between two surface forms (in this case, output-output correspondence; Benua 1997) rather than input-output correspondence. Because of their connection to the P-map (discussed in §2.2 below), the *MAP constraints must have access to the perceptual similarity of the two forms in correspondence; this notion seems less well defined for correspondences involving abstract input forms (see Zuraw 2013 for discussion). Second, I treat the *MAP constraints as symmetric here, meaning that *MAP(p, β) and *MAP(β, p) are considered equivalent. Having asymmetric versions of the constraints is not critical for the present study, though exploring this difference would be an interesting direction for future work.

It is worth noting that *MAP constraints are not necessarily inconsistent with the traditional correspondence constraints proposed by McCarthy & Prince (1995); some of the most widely used correspondence constraints can be treated as special cases of *MAP constraints. For instance, *MAP([αvoice], [–αvoice]) would be violated whenever there is a change in voicing, making it identical to IDENT(voice). Not all traditional correspondence constraints, however, can be translated straightforwardly into *MAP constraints (see Zuraw 2013 for a discussion).

Adopting *MAP constraints, we see that OT straightforwardly allows saltation, without requiring any other modifications. The solution for the Campidanian Sardinian case, where /VpV/ ⟶ [VβV] but /VbV/ remains unchanged, is shown in the tableaux in (5). The markedness constraints *V[−cont]V and *V[−voice]V are ranked above *MAP(p, β) so that underlying /VpV/ will surface as [VβV]. *MAP(b, β) can then be ranked above *V[−cont]V so that underlying /VbV/ will surface as [VbV], thus avoiding [b ~ β] alternations in cases like [**b**ĩu] 'wine', [sːu **b**ĩu] 'the wine'.[3]

[3] Hayes & White (2015) consider two additional ways that the constraint set could be modified so that saltation is possible in OT, namely Local Constraint Conjunction (Smolensky 2006) and Comparative Markedness (McCarthy 2003). The Local Constraint Conjunction approach to saltation requires conjoining a faithfulness constraint and a markedness constraint (see Lubowicz 2002, Ito & Mester 2003). The Comparative Markedness approach depends on markedness constraints that are violated only by marked structures present in the output form that were not present in the input form (i.e. 'new' violations; McCarthy 2003). Both accounts give markedness constraints the ability to apply only when forms are unfaithful, linking saltation to the concept of derived environment effects. Hayes & White (2015) argue that a *MAP approach, sufficiently constrained, is more restrictive than either Local Constraint Conjunction (specifically, the conjunction of faithfulness and markedness constraints) or Comparative Markedness, both of which predict highly implausible phonotactic patterns. For reasons of space, I will not repeat those arguments here.

(5) Deriving saltation in OT with *Map constraints

a. /VpV/ ⟶ [VβV]

/VpV/	*Map(b, β)	*V[−cont]V	*V[−voice]V	*Map(p, β)	*Map(p, b)
☞ VβV				*	
VbV		*!			*
VpV		*!	*		

b. /VbV/ ⟶ [VbV]

/VbV/	*Map(b, β)	*V[−cont]V	*V[−voice]V	*Map(p, β)	*Map(p, b)
VβV	*!				
☞ VbV		*			

However, having a theory that allows saltation is only half of the problem. Merely adding the *Map constraints to the theory is not sufficiently restrictive for two reasons. First, saltation appears to be a marked pattern. As mentioned above, cases of regular, stable saltation appear to be rare in the world's languages. Moreover, artificial language studies have shown that saltation is a dispreferred pattern during learning. For instance, White (2014) found that adults who learned potentially saltatory alternations (e.g. [p ~ v]) were biased to assume that phonetically intermediate sounds (e.g. [b]) also alternated, thereby avoiding the saltation. (These results are used to test the model's predictions below, and thus are discussed in more detail in §4.) Other studies have shown that 12-month-old infants have a similar bias against saltatory patterns when learning novel alternations, both in an artificial language (White & Sundara 2014) and in their native language (Sundara et al. 2013).

Second, and more generally, *Map constraints are much more powerful than traditional faithfulness constraints. Indeed, *Map constraints can do what traditional faithfulness constraints can do and much more. Thus we should be concerned about the implications of adding such a powerful tool to our phonological framework. For instance, if *Map constraints can be freely ranked, then why is there not a preponderance of completely arbitrary patterns across the world's languages?[4] These observations suggest that the *Map constraints are not freely ranked; rather, the theory must be constrained such that marked patterns, like saltation, are dispreferred.

[4] There are, of course, cases of unnatural/arbitrary patterns that occur in languages (e.g. Hellberg 1978, Anderson 1981), and such cases may not be as uncommon as we often assume (e.g. Mielke 2008). These patterns must be

To summarize, without expanding the constraint set beyond traditional feature-based faithfulness constraints, it is not possible to generate saltation. *MAP constraints make it possible for large changes to be preferred over small changes, which is essential for deriving saltation. However, given the apparent rarity of saltation, along with experimental evidence suggesting that saltation is dispreferred and general concerns about the powerful nature of the *MAP constraints, the theory needs to be sufficiently constrained.

The next section describes the proposed bias: a substantive bias based on the P-map (Steriade 2009 [2001]).

2.2. CONSTRAINING THE THEORY: THE P-MAP. Steriade (2009 [2001]) proposes that learners are equipped with a P-map (perceptibility map), representing knowledge that speakers have about the relative perceptual distance between any two pairs of sounds in a given phonological context. Under Steriade's account, the P-map is used as the basis for establishing *a priori* rankings of the correspondence constraints in a way that enforces a minimal modification bias in learners: phonological alternations are preferred if they result in minimal perceptual changes. Steriade supports her proposal with evidence from language typology, suggesting that markedness violations tend to be resolved in ways that require perceptually minimal modifications. The P-map has since been used to explain phenomena in a variety of languages (e.g. devoicing in Japanese, Kawahara 2006; cluster splittability in Tagalog and several other languages, Fleischhacker 2005, Zuraw 2007).

Saltation represents a gross violation of the P-map because rather than being a case of minimal modification, it is a clear case of EXCESSIVE modification. To have a saltatory alternation, a language must tolerate a large phonetic change (e.g. [p] ⟶ [β]) while not allowing a smaller change (e.g. [b] ⟶ [β]). The change from [p] to [β], for instance, is excessive given that [b] is legal in the given context and would require a less extreme change.

Zuraw (2007, 2013) proposes that the perceptual knowledge encoded in the P-map is translated into *a priori* rankings for *MAP constraints. Thus, *MAP constraints penalizing

learnable at some level. Indeed, saltation is an example of a pattern that I argue is dispreferred but learnable. It is therefore desirable that our phonological theory have the ability to account for arbitrary but learnable patterns when they do arise.

correspondences between perceptually dissimilar sounds (in a given context) are ranked higher than constraints penalizing correspondences between similar sounds. This *a priori* ranking represents the default ranking for the *MAP constraints, but the default hierarchy can be overturned if contradicted by sufficient evidence in the learner's input.

I adopt Zuraw's proposal that the *MAP constraints are constrained by a substantive bias based on the P-map. However, I follow Wilson (2006) in implementing this bias computationally via the prior probability distribution (henceforth just 'prior') of a maximum entropy (MaxEnt) model.[5] Instead of the constraints having an *a priori* strict ranking, they are assigned individual *a priori* preferred weights. Intuitively, these weights bias the learner to consider changes between similar sounds to be more likely than changes between dissimilar sounds, consistent with the principle of minimal modification inherent in the P-map theory. Based on the difference in relative similarity between the sounds involved, *MAP(p, β) will have a higher preferred weight than *MAP(b, β). However, as we saw in (5), to have saltation it must be possible to subvert this preferred hierarchy; that is, it must be possible for MAP(b, β) to attain a higher weight than *MAP(p, β) despite the P-map bias. A virtue of the MaxEnt learning framework is that constraint weights can gradually shift away from their prior values during the learning process as the model receives input data.

The next section discusses the details of the MaxEnt model.

3. IMPLEMENTING THE MODEL.

3.1. OVERVIEW OF MAXIMUM ENTROPY MODELS. MaxEnt models represent a general type of statistical model that has been used in a wide range of fields. They were first used to model phonological grammars by Goldwater and Johnson (2003) and have since been used in several other studies (e.g. Wilson 2006, Hayes & Wilson 2008, Hayes et al. 2009, Martin 2011, Hayes et al. 2012, Pater et al. 2012). As implemented here, the framework has a clear connection to OT (Prince & Smolensky 2004 [1993]) and Harmonic Grammar (Legendre et al. 1990, Smolensky & Legendre 2006), as has been discussed by others (Eisner 2000, Johnson 2002, Goldwater & Johnson 2003, Hayes et al. 2009, Culbertson et al. 2013). Since the model is essentially a

[5] The details of Wilson's implementation and the implementation here, however, are quite different; see §8.

probabilistic instantiation of Harmonic Grammar, the framework is sometimes called Probabilistic Harmonic Grammar (PHG; e.g. Culbertson et al. 2013).

For each input form, the MaxEnt model generates a probability distribution over the set of candidate output forms based on their violations of a set of weighted constraints. Specifically, for some input *x*, it assigns a probability to each output candidate *y* as in (6).

(6) $$\Pr(y \mid x) = \frac{\exp(-\sum_{i=1}^{m} w_i C_i(y,x))}{Z}, \text{ where } Z = \sum_{y \in Y(x)} \exp(-\sum_{i=1}^{m} w_i C_i(y,x))$$

Based on (6), the method for calculating the probability of an output candidate *y* for an input *x* can thus be described as follows. First, for each constraint, multiply the weight w_i of that constraint by the number of times the input/output pair violates the constraint, $C_i(y, x)$, and then sum over all constraints $C_1 \ldots C_m$. This summed value, $\sum w_i C_i(y, x)$, is comparable to the Harmony value from Harmonic Grammar (Legendre et al. 1990, Smolensky & Legendre 2006, Pater 2009) and has also been called a Penalty score (Hayes & Wilson 2008). Raise *e* to the negative Penalty score and finally divide the result by the sum over all possible output candidates (all *y* in the set Y(*x*)) for that input *x*. The sum over all output candidates is typically represented as *Z*.

The model is assumed to have a component comparable to GEN (i.e. Y(*x*) in the formula in (6)), which generates the set of output candidates for a given input form. The set of candidates is then evaluated on the basis of the grammar. In classical OT, candidates are evaluated based on a strict ranking of the constraints, such that one candidate is judged the winner if it is preferred by (i.e. has fewer violations of) the highest ranked constraint in the hierarchy (Prince & Smolensky 2004 [1993]). Constraints lower in the hierarchy have an influence on the outcome only if all higher ranked constraints have no preference among the candidates (i.e. only if all candidates have the same number of violations for each of the higher ranked constraints). Only one candidate (or set of candidates with identical violation profiles) is declared the winner in classical OT; all other candidates are losers. Thus, classical OT is not an effective framework for modeling variation or gradient outcomes, in which a single input may have multiple possible outputs.

The EVAL component of the MaxEnt model generates a probability distribution over all possible candidates for a given input, and unlike classical OT, the total probability may be divided unequally across different candidates. If the constraint weights are sufficiently different from one another,[6] the MaxEnt grammar will mimic the strict constraint rankings from classical OT, such that only one effective winner will emerge with a probability very close to 1.[7] In fact, it is possible to generate a MaxEnt simulation for any categorical outcome analyzed with classical OT as long as there is a finite limit on the number of constraint violations (Johnson 2002). However, if the constraint weights are similar to each other, then multiple candidates will be assigned probabilities that are not vanishingly small. In such cases, the model predicts that there will be variation in which output will be chosen; and crucially for the current study, the predicted probabilities for the output candidates can be compared to real data collected from a corpus or an experiment.

MaxEnt is one of several constraint-based models that have been proposed for handling phonological variation (e.g. see Anttila 1997, Ross 1996, Nagy & Reynolds 1997, Hayes & MacEachern 1998, Boersma & Hayes 2001, Boersma & Pater 2008). MaxEnt models, however, have two characteristics that distinguish them from many other approaches to modeling variation. First, MaxEnt models involve summing the violations of multiple weighted constraints (as in Harmonic Grammar; Legendre et al. 1990, Boersma & Pater 2008) rather than following a strict ranking hierarchy. As a result, MaxEnt models have the property of cumulative constraint interaction, often called 'ganging', whereby multiple violations of lower constraints can add up to overcome a violation of a constraint with a greater weight (Hayes & Wilson 2008, Pater 2009). Second, MaxEnt models are particularly attractive because they are associated with a learning algorithm (Berger et al. 1996) that provably converges on the objectively optimal grammar, which in MaxEnt is defined as the set of constraint weights that maximizes the probability of the training data (taking into account the prior, see below). By comparison, it has been shown that the Gradual Learning Algorithm (GLA; Boersma & Hayes 2001) sometimes fails to converge on a grammar, even when a grammar exists that could, in principle, account for the data (Pater 2008).[8]

[6] To achieve this, constraint weights need to be spaced at roughly exponential increments, see Johnson 2002 and Goldwater & Johnson 2003.

[7] The probability can never actually reach 1 because other candidates must receive some probability, even if vanishingly small; in other words, the numerator in (6) can never reach 0 for any given candidate.

[8] A modification to the GLA by Magri (2012) allows it to successfully handle the case put forth by Pater (2008). However, in unpublished work, Bruce Hayes (personal communication) has discovered another case in which the GLA fails to converge, even when Magri's modification is used.

3.2. LEARNING THE WEIGHTS. Given a set of constraints and the observed data, the learning problem for the MaxEnt model is to find the weights that maximize the probability of the observed data (thereby minimizing the probability of unobserved data). The probability of the observed data, *D*, is calculated by taking the product of the model-predicted conditional probabilities of each output observed during training given its input, $\{(y_1 \mid x_1) \ldots (y_n \mid x_n)\}$, as in (7).

(7) $$\Pr(D) = \prod_{j=1}^{n} \Pr(y_j \mid x_j)$$

This calculation is computed on the basis of observed tokens, so 100 examples of the input/output pair (b | p) during training will have a greater effect on the model than only one example. Because probabilities are being multiplied, the Pr(*D*) calculated in (7) becomes extremely small, so in practice the calculation is done by taking the sum of the log probabilities of each output given its input, as in (8).

(8) $$\log \Pr(D) = \sum_{j=1}^{n} \log \Pr(y_j \mid x_j)$$ (equivalent to log (7))

The model also takes into account a regularizing bias term, the 'prior', during learning. The prior term is a Gaussian distribution over each constraint weight, defined in terms of a mean, μ, and a standard deviation, σ, as in (9).

(9) $$\sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2}$$

The result of (9) is subtracted from the probability calculated in (6). The μ for each constraint acts as its *a priori* 'preferred' weight, which is subtracted from the constraint's learned weight, *w*; the difference in actual and preferred weight is then squared. Thus, as constraints vary more from their μ, the penalty imposed by the prior increases. The value of σ^2 determines how tightly each constraint's weight is constrained to its μ. Because it is in the denominator, lower values of σ^2 result in a greater penalty for weights that vary from their μ. Having low values of σ^2 therefore means that more data are required to move the weights away from μ during learning. Higher values of σ^2 mean that the weights have more freedom to vary from their μ. In sum, the prior acts as a penalty that increases as constraint weights diverge from their *a priori* preferred weights.

When the prior is uniform across all constraints (and σ^2 is not set very high), the model prefers grammars in which weight is distributed among each of the constraints, and ample amounts of data are needed for constraints to reach relatively extreme weights. For this reason, Gaussian priors are commonly used in MaxEnt models as a way to prevent overfitting (discussed, e.g., by Goldwater & Johnson 2003). In the model presented here, constraints may each receive a different μ, so the prior also serves as a means of implementing a substantive learning bias (see §3.3).

With the inclusion of a prior, the goal of learning is to choose the set of constraint weights that maximizes the objective function in (10), in which the prior term in (9) is subtracted from the log probability of the observed data (the function in (8)).

(10) $$[\sum_{j=1}^{n} \log \Pr(y_j \mid x_j)] - [\sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2}]$$

The search space of log likelihoods is provably convex, meaning that there is always one objective set of weights that will maximize the function in (10), and this set of weights can be found using any standard optimization strategy (Berger et al. 1996). Here, the model was implemented using the MaxEnt Grammar Tool,[9] which uses the Conjugate Gradient algorithm (Press et al. 1992) to find the weights during learning.

3.3. SETTING THE PRIOR. I implemented three versions of the model: one with a substantive bias based on the P-map, one in which all constraints have a preferred weight of 0, and one with no substantive bias, but a general preference for non-alternation. The latter two models will serve

[9] Software developed by Colin Wilson and Ben George, made available for public use by Bruce Hayes at http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/.

as controls for the substantively biased model. The basic architecture of the three models is the same. The only difference is how the prior is set in each model.

SUBSTANTIVELY BIASED MODEL. The substantively biased version of the model was implemented by assigning each *MAP constraint an individual preferred weight (μ) based on the perceptual similarity of the pair of sounds specified by the constraint (for a similar use of prior μ values as a way to implement a substantive bias in the domain of syntax, see Culbertson et al. 2013). For this implementation, two main issues needed to be resolved: (1) how to define perceptual similarity, and (2) how to generate the μ for each constraint based on the measure of similarity chosen. These issues are discussed in turn.

First, determining how to define and measure perceptual similarity is not trivial. In reality, listeners probably take many factors into account when making such judgments (e.g. see Steriade 2009 [2001], Mielke 2012, Cristia et al. 2013). Here, I use confusability as an approximation of perceptual similarity, where the confusability of two speech sounds is determined according to the results of standard identification experiments (e.g. Miller & Nicely 1955, Singh & Black 1966, Wang & Bilger 1973, Cutler et al. 2004). In these experiments, participants listen to recordings of speech sounds (with or without noise) and identify the sounds that they hear in some target location. A confusion matrix can then be calculated based on the responses recorded for each sound. Even if somewhat coarse, confusability is a straightforward way of approximating perceptual similarity, and it works well for the purposes of this study.

I used confusion data reported in Wang and Bilger 1973, whose participants were native English speakers. Specifically, I summed the values from Tables 2 and 3 of Wang and Bilger 1973, where the target consonants were placed in CV (W&B: Table 2) and VC (W&B: Table 3) contexts. The stimuli were presented in noise, and the values from these two tables represent the summed values across all signal-to-noise ratios. Reasons for using these values, and potential implications of doing so, are discussed further in §6.2.

The second consideration was how to go from the confusion probabilities (i.e. the results from perception experiments) to the preferred weights for the prior. To accomplish this, the confusion values were entered into a separate MaxEnt model intended only to generate the prior weights. Intuitively, one can think of this model as representing the learner's experience perceiving speech sounds. For reference, the confusion values given to the model are provided in

Table 1. Each relevant *MAP constraint was included in the model, and violations were marked whenever the two sounds listed in the constraint were confused for one another. For instance, a violation was marked for *MAP(p, v) when [p] was confused for [v], or *vice versa*.[10]

	Responses					Responses			
Stimulus	p	b	f	v	Stimulus	t	d	θ	ð
p	1844	54	159	26	t	1765	107	92	26
b	206	1331	241	408	d	91	1640	75	193
f	601	161	1202	93	θ	267	118	712	135
v	51	386	127	1428	ð	44	371	125	680

TABLE 1. Confusion values for the combined CV and VC contexts from Wang & Bilger 1973 (Tables 2 and 3), which were used to generate the prior. Only the sound pairs relevant for the current study are shown here.

The *MAP constraints then received weights based on how often the two sounds named in the constraint were confused for each other in the confusion experiment. The resulting weights are provided in Table 2. Sounds that are very confusable, and thus assumed to be highly similar, resulted in low weights whereas sounds that are dissimilar resulted in more substantial weights. For instance, [b] and [v] are very similar to each other, so *MAP(b, v) received a small weight of 1.30. On the other hand, [p] and [v] are quite dissimilar, so *MAP(p, v) received a greater weight of 3.65. These weights were entered directly into the primary learning model's prior as the preferred weights (μ) for the *MAP constraints. It is preferable to derive the prior weights directly from the confusion data in a systematic way, as was done here, rather than 'cherry picking' the set of prior weights that results in the best performance. The systematic approach taken here allows us to draw better conclusions about how the success (or failure) of the model relates to the relationship between perceptual similarity and the learning process.

The σ^2 was set to 0.6 for every constraint, which was the value that maximized the fit of the model's predictions to the experimental results. Other values of σ^2 are considered in §6.1.

[10] For the prior, μ was set to 0 and σ^2 was set to 10,000. This value of σ^2 is sufficiently high that the prior had very little influence; the weights were essentially free to be whatever they needed to be in order to best predict the confusion probabilities in the input data.

Labial sounds		Coronal sounds	
Constraint	Prior weight (μ)	Constraint	Prior weight (μ)
*MAP(p, v)	3.65	*MAP(t, ð)	3.56
*MAP(f, v)	2.56	*MAP(θ, ð)	1.91
*MAP(p, b)	2.44	*MAP(t, d)	2.73
*MAP(f, b)	1.96	*MAP(θ, d)	2.49
*MAP(p, f)	1.34	*MAP(t, θ)	1.94
*MAP(b, v)	1.30	*MAP(d, ð)	1.40

TABLE 2. Prior weights (μ) for *MAP constraints in the substantively biased model, based on confusion data from Wang & Bilger (1973).

UNBIASED MODEL. The second version of the model, which I will call the 'unbiased' model, had no substantive bias. The unbiased model had a 'flat' prior: every constraint had the same μ (set to 0, where 0 means the constraint has no effect on the outcome) and σ^2 (set to 0.6, i.e. the same value as in the biased model).[11] It was otherwise identical to the substantively biased model.

ANTI-ALTERNATION MODEL. In the unbiased model, the μ for every constraint was set to 0, but in the substantively bias model, each *MAP constraint had a non-zero weight. Thus, the unbiased model may not be the fairest comparison because it differed from the substantively biased model on two accounts: (1) not having a substantive bias, and (2) having faithfulness biased towards 0 rather than some positive weight. When we test the models below, we will see that this difference is indeed important.

To address this issue, I implemented a third model, called the 'anti-alternation model', in which every *MAP constraint was assigned a prior weight of 2.27. This value is the mean of all the *MAP prior weights in the substantively biased model. The mean of the prior weights in the biased model was chosen in order to give the anti-alternation model the best chance of succeeding. In all other ways, the model was identical to the other two models. The anti-

[11] In principle, a better comparison might involve fitting the unbiased model (and the anti-alternation model below) with the σ^2 that maximizes its own performance rather using the σ^2 that maximizes the performance of the substantively biased model. In practice, this turns out not to matter much; the unbiased model and the anti-alternation model never approach the level of success that the substantively biased model achieves, regardless of which value of σ^2 is used; see §6.1.

alternation model is similar to the substantively biased model in that all of the *MAP constraints have non-zero weights; but unlike the substantively biased model, those weights do not vary between the constraints according to perceptual similarity.

Because the *MAP constraints are best conceptualized as output-output faithfulness constraints (Benua 1997) or paradigm uniformity constraints (Hayes 1997, Steriade 2000), having a non-zero prior for these constraints is akin to having a default preference to avoid any alternation at all. I return to this point in §9.

4. THE EXPERIMENTAL DATA TO BE MODELED. The data to be modeled come from two artificial language learning experiments reported in White 2014. Here, I provide only an overview of the experiments and the results; the reader is referred to the original paper for greater detail on the methodology and the statistical analysis.

4.1. EXPERIMENT 1. In Experiment 1, English-speaking participants were divided into two conditions. The Potentially Saltatory condition learned two potentially saltatory alternations, [p ~ v] and [t ~ ð], during training. Participants heard a singular nonce form (e.g. [luˈman]) followed by the corresponding plural form with an additional plural suffix *-i* (e.g. [luˈmani]); these forms were presented with singular and plural pictures, respectively. Half of the trials involved filler sounds with no change in the final consonant (e.g. [luˈman] … [luˈmani]), but the other half included final consonants demonstrating the crucial alternations (e.g. [kaˈmap] … [kaˈmavi]).

After training, participants were tested using a forced-choice task. They were presented with novel singular words ending in trained sounds (e.g. [suˈlap]) and were asked to choose between two plural options: a non-changing plural form ([suˈlapi]) and a changing plural form ([suˈlavi]). Crucially, they were also tested on singular words ending in the phonetically intermediate sounds [b, f, d, θ], which were absent from training. The second group of participants (Control condition) had a similar experience; however, they learned alternations that were not potentially saltatory ([b ~ v] and [d ~ ð]) and were tested on a set of untrained sounds, [p, f, t, θ], that were not phonetically intermediate between the alternating sounds in their condition.

The crucial results are presented in Figure 2 (left side). As shown by the dark grey bars, participants in both conditions successfully learned the alternations presented during training; they correctly changed these sounds at test. Looking at the untrained sounds (light grey bars), we see that participants in the Potentially Saltatory condition frequently generalized to the intermediate sounds [b, f, d, θ], even though there was no evidence for such changes during training. In contrast, participants in the Control condition generalized to untrained sounds much less frequently compared to participants in the Potentially Saltatory condition. The greater generalization to untrained sounds in the Potentially Saltatory condition (where untrained sounds were phonetically intermediate between the alternating sounds) compared to the Control condition (where untrained sounds were not intermediate) suggests that participants were biased to avoid saltation. By changing untrained sounds in the Potentially Saltatory condition, participants could avoid a saltatory system; the trained alternations in the Control condition, on the other hand, were non-saltatory regardless of whether participants changed the untrained sounds.

Finally, in the Potentially Saltatory condition, the results revealed a statistically significant preference for changing voiced stops ([b, d]; third and fourth bars) compared to changing voiceless fricatives ([f, θ]; fifth and sixth bars). We will return to this difference when we compare the model predictions to the experimental results in §5.2.

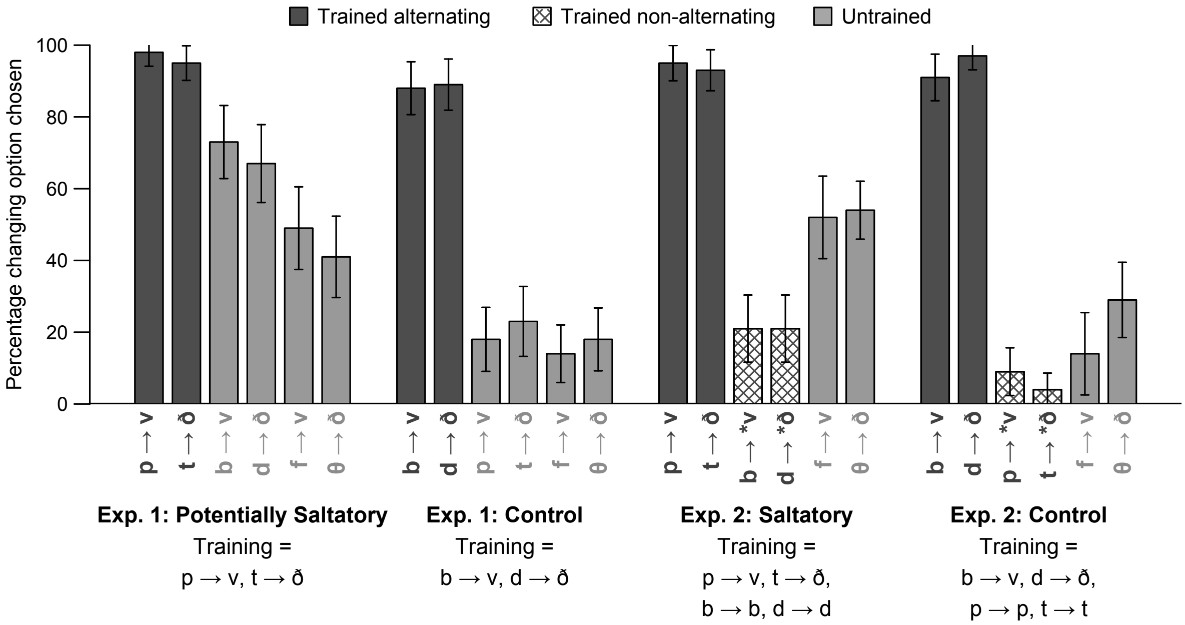

FIGURE 2. Experimental results from White 2014 for Experiment 1 (left side) and Experiment 2 (right side), according to condition. Bars indicate percentage of trials in which the changing option was chosen according to the final sound of the singular form. Errors bars show standard error of the mean.

4.2. EXPERIMENT 2. Participants in Experiment 2 were likewise divided into two conditions. In the Saltatory condition, participants learned the same alternations as in the Potentially Saltatory condition of Experiment 1 (i.e. [p ~ v] and [t ~ ð]), but they additionally received explicit training that the intermediate voiced stops [b, d] did not alternate. Thus, because these participants were learning alternations with non-alternating intermediate sounds, they were tasked with learning an explicitly saltatory system. Participants in the Control condition learned the same alternations as those in the Control condition of Experiment 1 (i.e. [b ~ v] and [d ~ ð]), but they were additionally trained that the non-intermediate voiceless stops [p, t] did not alternate. The results are shown in Figure 2 (right side).

Participants were once again successful at learning the alternations presented during training (indicated by the dark grey bars). The crucial results in Experiment 2 were the stops that participants were trained should not change (indicated by the cross-hatched bars). Participants in the Control condition only rarely changed the non-intermediate [p, t] in error (7% of trials on average), but participants in the Saltatory condition were more likely to change intermediate [b, d] in error (21% of trials), in spite of their training. The results once again suggest an anti-saltation bias: participants in the Saltatory condition found it difficult to learn that intermediate sounds did not alternate. Finally, the untrained fricatives in Experiment 2 (light grey bars) replicated the findings from Experiment 1; these sounds were changed more often in the Saltatory condition than in the Control condition.

In the next section, we test the three models (i.e., the substantively biased, unbiased, and anti-alternation models) by comparing their predictions to the experimental results.

5. TESTING THE MODELS. To assess the MaxEnt models' predictions, the models were provided the same training data as the experimental participants, summarized in Table 3. Recall that each model was also provided with a set of constraints (two markedness constraints and a set of relevant *MAP constraints), and one of three priors, as discussed in §3. Given the training data and the prior, each model learned a grammar (i.e. a set of weights for each constraint; see §3.2), and was then tested on the same test items seen by the experimental participants. The resulting model predictions (discussed in the following sections) were then compared to the aggregate experimental results to determine how well each model approximated human performance.

Experiment 1		Experiment 2	
Potentially Saltatory condition	Control condition	Saltatory condition	Control condition
18 p ⟶ v	18 b ⟶ v	18 p ⟶ v	18 b ⟶ v
18 t ⟶ ð	18 d ⟶ ð	18 t ⟶ ð	18 d ⟶ ð
		9 b ⟶ b	9 p ⟶ p
		9 d ⟶ d	9 t ⟶ t

TABLE 3. Overview of training data for the MaxEnt model, based on the experiments in White 2014.

5.1. GENERATING THE MODEL PREDICTIONS. During learning, the model considered all obstruents within the same place of articulation as possible outputs for a given input. For instance, for input /p/, the model considered the set {[p], [b], [f], [v]} as possible outputs. In the observed training data, however, each input had only one possible output because there was no free variation in the experimental training data (e.g. in Experiment 1 singular-final [p] or [b], depending on condition, changed to [v] 100% of the time in the plural). Thus, during learning the model was trying to account for the fact that the observed output was the winner and the other three possible outputs were losers. At test, the model only considered the relative probability of two possible outputs – that is, the two outputs that the experimental participants considered in the forced-choice task. The goal was to put the model and the participants in the same situation: during training, neither the participants nor the model knew the format of the test phase, so they had to consider all possible outputs for each input. But at test, the model and participants alike were forced to choose between only two possible outputs.

As an example, consider how the predictions were generated for the substantively biased model in Experiment 1. Table 4 shows how the constraint weights changed from their prior weights as a result of the training data in Experiment 1.

In the Potentially Saltatory condition (training = p ⟶ v; t ⟶ ð), we see that both markedness constraints, *V[–voice]V and *V[–cont]V, pick up weights so that voiceless stops will be changed to voiced fricatives. Likewise, the weights for *MAP(p, v) and *MAP(t, ð) are substantially reduced because the training data provide evidence that those mappings indeed occur. Other *MAP constraints involving [p] and [t] (i.e. *MAP(p, b), *MAP(t, d), *MAP(p, f), *MAP(t, θ)) have modest increases in their weights because they all play a role (during learning though not at test, where only two outcomes are possible) in ensuring that [p] and [t] are mapped to [v] and [ð], respectively, rather than to some other sound ([b], [f], [d], or [θ]). Weights for the remaining *MAP constraints remain at their prior weights because they have no effect on the [p] ⟶ [v] or [t] ⟶ [ð] outcomes.

In the Control condition (training = b ⟶ v; d ⟶ ð), the markedness constraint *V[–cont]V gets a substantial increase in weight to motivate spirantization. The markedness constraint *V[–voice]V receives a small increase in weight due only to its limited role in preventing [b] and [d] from changing into [p] and [t], respectively. *MAP(b, v) and *MAP(d, ð) have substantially reduced weights (almost to 0) because the training data indicate that such mappings are allowed. Once again, other *MAP constraints involving [b] or [d] receive modest increases in their weights so that [b] and [d] will be mapped to [v] and [ð], respectively, rather than to some other sound.

The weights of the remaining *MAP constraints, which have no effect on the outcomes of the training data, remain at their prior values.

		Post-learning weight	
Constraint	Prior weight	Potentially Saltatory condition	Control condition
*V[–voice]V	0	2.20	0.57
*V[–cont]V	0	1.86	1.80
*MAP(p, v)	3.65	2.17	3.65
*MAP(t, ð)	3.56	2.22	3.56
*MAP(p, b)	2.44	2.77	2.48
*MAP(t, d)	2.73	3.02	2.76
*MAP(p, f)	1.34	1.90	1.34
*MAP(t, θ)	1.94	2.34	1.94
*MAP(b, v)	1.30	1.30	0.15
*MAP(d, ð)	1.40	1.40	0.25
*MAP(f, v)	2.56	2.56	2.56
*MAP(θ, ð)	1.91	1.91	1.91
*MAP(b, f)	1.96	1.96	2.25
*MAP(d, θ)	2.49	2.49	2.70

TABLE 4. Prior constraint weights and post-learning weights (substantively biased model) in the Potentially Saltatory and Control conditions of Experiment 1.

From the post-learning weights, the model calculates the predicted probability of each output candidate at test, given each input. These probabilities are calculated as described in §3.1. Representing the calculations in an OT-style tableau highlights the MaxEnt grammar's similarity to OT and Harmonic Grammar. A couple of examples are given in (11) for inputs /VpV/ and /VbV/ in the Potentially Saltatory condition of Experiment 1. Constraint weights are taken from Table 4.

(11) Calculating predicted probabilities in tableaux

a. Input /VpV/ in Experiment 1, Potentially Saltatory condition

/VpV/	*V[−voice]V 2.20	*MAP(p, v) 2.17	*V[−cont]V 1.86	*MAP(b, v) 1.30	Penalty score	$e^{(-\text{penalty})}$	Predicted prob.
VvV		1			2.17	.1142	.87
VpV	1		1		4.06	.0172	.13

b. Input /VbV/ in Experiment 1, Potentially Saltatory condition

/VbV/	*V[−voice]V 2.20	*MAP(p, v) 2.17	*V[−cont]V 1.86	*MAP(b, v) 1.30	Penalty score	$e^{(-\text{penalty})}$	Predicted prob.
VvV				1	1.30	.2725	.64
VbV			1		1.86	.1557	.36

The predicted probabilities of each output candidate can then be compared to the experimental results. In the following sections, we will take a detailed look at the predictions of each of the three models (substantively biased, unbiased, and anti-alternation) for Experiment 1 and Experiment 2. The predictions for each model are shown in Figure 3, superimposed on the experimental results for easy comparison.[12] For reference, the prior and post-learning weights for each model (like those presented in Table 4) are provided in the Supplementary Materials, along with greater discussion of why the weights changed as they did.

5.2. EXPERIMENT 1. Based on the results from Experiment 1 (§4.1), there are three main patterns that the learning model should be able to capture. First, participants were successful at learning the alternations that they were trained on (i.e. [p ~ v] and [t ~ ð] in the Potentially Saltatory condition; [b ~ v] and [d ~ ð] in the Control condition). Second, participants generalized to untrained sounds more often in the Potentially Saltatory condition, where the untrained sounds were phonetically intermediate, than in the Control condition where they were not intermediate. Finally, in the Potentially Saltatory condition, there was a significant preference to change the voiced stops [b, d] compared to the voiceless fricatives [f, θ], even though neither group was presented during training.

Looking at the predictions for Experiment 1 (Figure 3, left side), we see that each of the models predicts a comparably high rate of success on the alternations that were presented during training (indicated by the dark grey bars). The more interesting comparison comes from looking at the models' predictions for the untrained sounds (light grey bars). The three models will be considered in turn.

[12] I would like to thank Jennifer Culbertson for kindly sharing the R code that was used (with modification) to make the plots in Figure 3, which are based on similar plots found in Culbertson et al. 2013.

SUBSTANTIVELY BIASED MODEL. The substantively biased model correctly predicts greater generalization to untrained sounds in the Potentially Saltatory condition than in the Control condition. In the Potentially Saltatory condition, the trained alternations ([p] ⟶ [v]; [t] ⟶ [ð]) motivate an increase in the weights of both markedness constraints, *V[−cont]V and *V[−voice]V. These two markedness constraints also motivate changing the intermediate sounds [b, d] and [f, θ], respectively. The weights of the *MAP constraints protecting the intermediate sounds from changing remain at their fairly low prior weights since these sounds did not appear during training. Therefore, we see a large amount of generalization to intermediate sounds.

In the Control condition, the model predicts much less generalization to untrained sounds. The trained alternations (b ⟶ v; d ⟶ ð) cause a large increase in the weight of *V[−cont]V, but only a tiny increase for *V[−voice]V. Because *V[−voice]V has a low weight, we see little generalization to [f] ⟶ [v] or [θ] ⟶ [ð]. Moreover, because *MAP(p, v) and *MAP(t, ð) have large prior weights, we see little generalization to [p] ⟶ [v] or [t] ⟶ [ð].

It is worth emphasizing that the substantively biased prior plays a key role in causing the saltation avoidance pattern in the model's predictions. In the Potentially Saltatory condition, the high prior weights of *MAP(p, v) and *MAP(t, ð) mean that the two markedness constraints must receive substantial weights in order for the trained alternations to be learned. This, coupled with the relatively low prior weights of the *MAP constraints protecting the intermediate sounds, results in a large amount of generalization. In the Control condition, the high prior weights of *MAP(p, v) and *MAP(t, ð) result in little generalization to those sounds. The consequence is asymmetric generalization that is consistent with the P-map: alternation between dissimilar sounds is generalized to similar sounds, but not *vice versa*.

Finally, the model also accounts for the preference observed in the Potentially Saltatory condition for spirantizing intermediate stops (i.e. [b ~ v] and [d ~ ð]) compared to voicing intermediate fricatives ([f ~ v] and [θ ~ ð]). This difference falls out directly from the P-map prior. Taking the labials as an example: because [b] and [v] are more perceptually similar than [f] and [v], *MAP(b, v) has a lower prior weight than *MAP(f, v). As a result, changing [f] to [v] garners a greater penalty than changing [b] to [v], leading to the difference observed.

UNBIASED MODEL. Looking at the unbiased model's predictions, we see first of all that the predictions are identical for labials and coronals, and that they are identical for voiced stops and voiceless fricatives in the Potentially Saltatory condition. Because the model does not have access to perceptual similarity, there is no *a priori* difference between, for instance, the alternations [b ~ v] and [f ~ v]. As a result, the model is unable to account for the significant difference between voiced stops and voiceless fricatives observed in the Potentially Saltatory condition.

Perhaps the most striking aspect of the predictions is how much the model overestimates the amount of generalization to untrained sounds in the Control condition. In fact, the model predicts more overall generalization to untrained sounds in the Control condition than in the Potentially Saltatory condition, opposite of the experimental results. This extreme amount of overgeneralization, however, is primarily due to the fact that the *MAP constraints all have a prior weight of 0 rather than the lack of substantive bias *per se*. Because the untrained sounds are absent in training, there is no reason to raise the weights of the *MAP constrains protecting these sounds, so the weights remain at 0. With this in mind, let us turn to the anti-alternation model, in which the *MAP constraints are all set with a non-zero prior weight, which is more comparable to the prior weights in the substantively biased model.

ANTI-ALTERNATION MODEL. The anti-alternation model also makes no distinction between labials and coronals; moreover, it is unable to account for the greater preference seen in the Potentially Saltatory condition for changing intermediate voiced stops compared to intermediate voiceless fricatives. Both outcomes follow from the fact that all of the *MAP constraints have the same prior weights.

The most critical problem of the model is that, like the unbiased model, it predicts the incorrect pattern of generalization for untrained stops. Specifically, it predicts more generalization in the Control condition (where untrained stops are not phonetically intermediate) than in the Potentially Saltatory condition (where they are phonetically intermediate), which is the opposite of the experimental results. This problem is once again due to the fact that the *MAP constraints have the same prior weights. Without the P-map bias, which assigns *MAP(p, v) a higher prior weight than *MAP(b, v), the model lacks the *a priori* knowledge that [b ~ v] is a more likely alternation than [p ~ v].

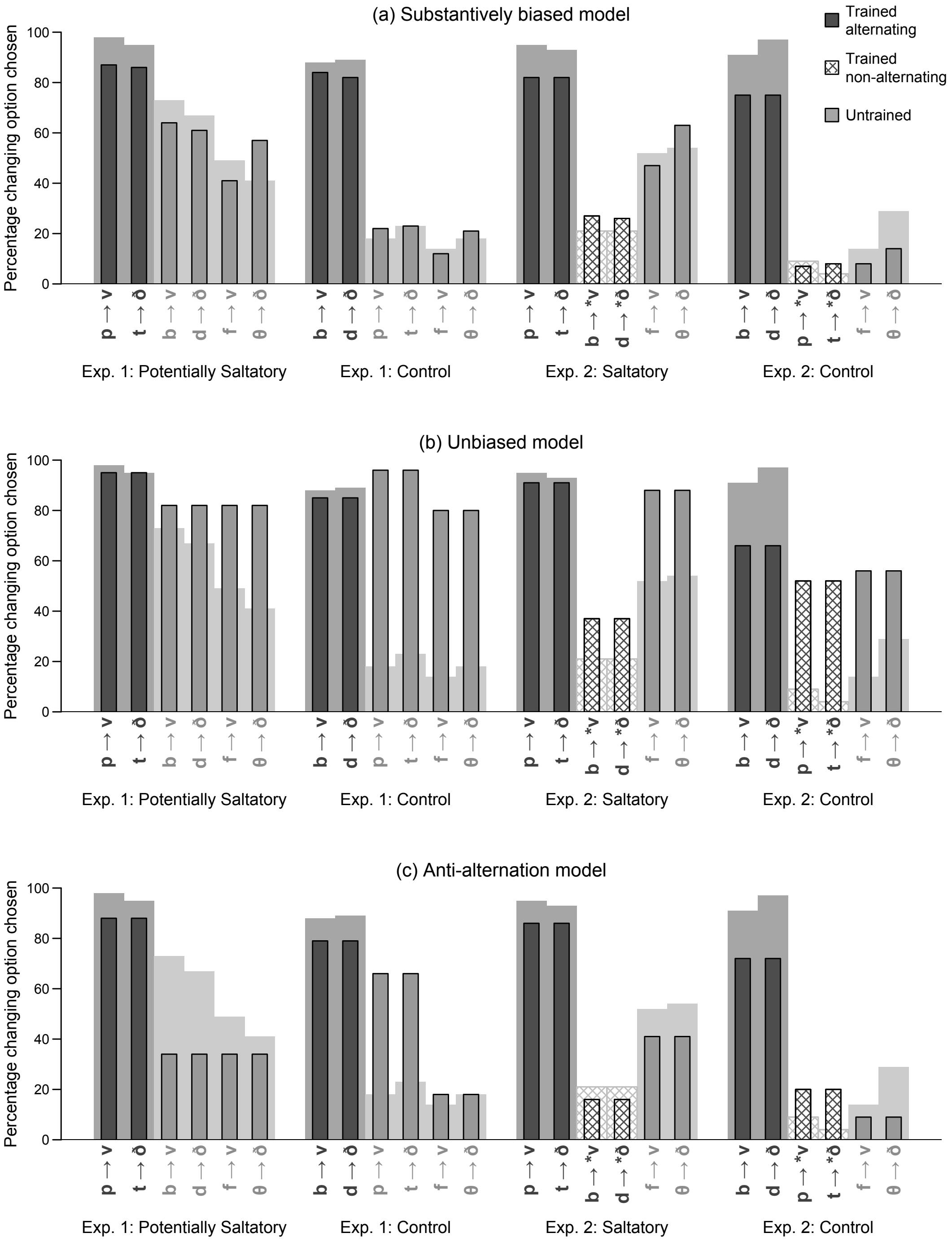

FIGURE 3. Predictions of the substantively biased, unbiased, and anti-alternation models superimposed on the experimental results, according to experiment and condition.

5.3. Experiment 2. Recall that in Experiment 2, participants were trained on the same alternations as in Experiment 1 (Saltatory condition: p ⟶ v and t ⟶ ð; Control condition: b ⟶ v and d ⟶ ð), but they were additionally trained that certain stops did not alternate (Saltatory condition: b ⟶ b and d ⟶ d; Control condition: p ⟶ p and t ⟶ t). Based on the experimental results (§4.2), there are two main patterns that the learning model should be able to capture. First, participants in the Saltatory condition were more likely than participants in the Control condition to change the trained non-alternating stops in error. Second, generalization to untrained fricatives was greater in the Saltatory condition than in the Control condition, replicating the basic effect from Experiment 1.

Looking at the predictions (Figure 3, right side), we see that the models have comparable predictions on the trained alternations (dark grey bars); therefore, we will focus on the trained non-alternating sounds (cross-hatched bars) and the untrained sounds (light grey bars). Each of the three models will again be considered in turn.

Substantively biased model. Consistent with the experimental results, the substantively biased model predicts that intermediate [b, d] in the Saltatory condition will be changed in error more frequently than [p, t] in the Control condition (Figure 3a, cross-hatched bars). This difference can be directly traced to the P-map bias implemented in the prior. In both conditions, the alternations learned during training (Saltatory condition: p ⟶ v and t ⟶ ð; Control condition: b ⟶ v and d ⟶ ð) lead to an increase in the weight of *V[−cont]V. However, the increased weight of *V[−cont]V also motivates spirantization of the non-alternating stops seen during training. The only way to protect these stops from changing is to raise the weights of the relevant *Map constraints even higher. In the Control condition, examples of non-changing [p] and [t] in training lead to increases in the weights of *Map(p, v) and *Map(t, ð); since these *Map already have high prior weights, it is easy to reach a sufficiently high weight to (mostly) overcome the rising weight of *V[−cont]V. In the Saltatory condition, cases of non-changing [b] and [d] in training also lead to increased weights for the relevant *Map constraints, *Map(b, v) and *Map(d, ð) in this case. However, because these *Map constraints have low prior weights, their weights do not reach a sufficiently high level to fully protect the intermediate stops, resulting in a large number of errors.

Additionally, as in Experiment 1, the model predicts a greater tendency to change untrained fricatives ([f] and [θ]) in the Saltatory condition than in the Control condition, consistent with the experimental results.[13]

UNBIASED MODEL. Unlike the substantively biased model, the unbiased model is not able to account for the basic anti-saltation effect. Specifically, it predicts more errors on voiceless stops in the Control condition than on voiced stops in the Saltatory condition, contrary to the experimental results (Figure 3b, cross-hatched bars). The model also overgeneralizes to untrained fricatives, just as it did for Experiment 1.

As in Experiment 1, the model's failures can be attributed to the fact that all *MAP constraints have a prior weight of 0. For the untrained fricatives, the weight of the relevant *MAP constraints never have a reason to rise above 0, leading to overgeneralization (i.e. the same problem as in Experiment 1). The reason that the model predicts too many errors for voiceless stops in the Control condition is somewhat subtler. Using the labials as an example, the only constraint that can motivate the [b] ⟶ [v] change found in the training data is *V[–cont]V; however, raising this constraint also causes [p] to change. The only way to protect [p] from changing is to increase the weight of *MAP(p, v). But because all of the constraints start at 0, the model is unable to raise the weight of *V[–cont]V enough to motivate the [b] ⟶ [v] alternation while also raising the weight of *MAP(p, v) enough to protect [p] from changing. *MAP(p, v) would need to be raised far above *V[–cont]V in order to protect [p]. By contrast, in the substantively biased model, *MAP(p, v) has a high prior weight, so achieving the necessary dispersion of the weights is possible.

[13] One point of concern is that the model consistently underestimates the performance on trained alternations. This is especially noticeable in the Control condition of Exp. 2, but it holds to a lesser extent in the other conditions as well. Note that the problem is not limited to the substantively biased model; rather, it is true of all three models. Two factors may be relevant here. First, White's (2014) experiments required participants to reach an 80% accuracy criterion on trained items before moving into the test phase, but this criterion is not taken into account in the model. This aspect of the experimental design may have artificially increased accuracy on trained alternations. Second, it is possible that learners have a regularization bias, biasing them towards applying processes 100% or 0% of the time, which is not taken into account by this model (but see Culbertson et al. 2013 for an implementation of a regularization bias in a MaxEnt model). These ideas are speculative at this point, and a more complete understanding of the issue is left for future research.

ANTI-ALTERNATION MODEL. The predictions of the anti-alternation model provide a good overall fit to the results from the Saltatory condition. However, like the unbiased model, the anti-alternation model fails to account for the crucial DIFFERENCE in the number of errors observed for trained non-alternating stops in the two conditions. Specifically, it predicts slightly more errors on the voiceless stops [p, t] in the Control condition than on the intermediate voiced stops [b, d] in the Saltatory condition (Figure 3c, cross-hatched bars). In the experimental results, there were very few such errors in the Control condition.

The reason for the anti-alternation model's failure once again stems from the model's lack of substantive bias. In the Control condition of Experiment 2, the model must account for [b] ⟶ [v] changes while also accounting for the fact that [p] is not spirantized. To do so, the weight of *V[–cont]V needs to be moderate, the weight of *MAP(b, v) needs to be low, and the weight of *MAP(p, v) needs to be high. In the substantively biased model, the prior weights make it easy to reach this arrangement: *MAP(b, v) already has a low prior weight and *MAP(p, v) already has a high prior weight. In the anti-alternation model, on the other hand, the prior weights for *MAP(b, v) and *MAP(p, v) are identical, which hinders the model's ability to reach the relative weighting necessary to account for the data.

5.4. OVERALL MODEL PERFORMANCE. As we observed in the previous sections, only the substantively biased model accounts for all of the qualitative differences found in the experimental results. How well do the models perform overall? We can consider the overall fit of the models by plotting the individual observations from Experiment 1 and Experiment 2 against the predictions of each model, as shown in Figure 4. Each point in the figure represents one observation, with the model prediction (x-axis) plotted against the mean experimental result across all participants (y-axis). For example, one point represents the mean percentage of [p] ⟶ [v] OBSERVED in the Control condition of Experiment 1 plotted against the PREDICTED percentage of [p] ⟶ [v] for that condition. Table 5 further shows the r^2 and log likelihood for each model, fitting the model predictions to the experimental results.

Overall, we see that the predictions of the substantively biased model produce an excellent fit to the observed experimental data, accounting for about 94% of the overall variance (r^2 = .94). Recall that the prior weights (μ) used in the substantively biased model were derived directly from the confusion data; these values were not fitted to produce the best outcome! In comparison, the unbiased model's predictions result in a very poor fit to the data (r^2 =.25). The anti-alternation model's predictions result in a better fit to the data (r^2 =.67) than the unbiased model, but its performance is nevertheless much worse than the substantively biased model. Log likelihood shows the same pattern: the substantively biased model predicts the greatest likelihood of the data, followed by the anti-alternation model, and then the unbiased model.

Looking at Figure 4, it appears that the anti-alternation model particularly falls short (relative to the substantively biased model) when it comes to differentiating the values in the middle part of the scale. Indeed, if we consider only the middle two-thirds of the experimental results (those with percentages falling between 10% and 90% changed), the r^2 value of the substantively biased model remains high (r^2 = .91), but the fit of the anti-alternation model declines drastically (r^2 = .36), suggesting that the distinction between these models is even greater for the subset of data with intermediate values.

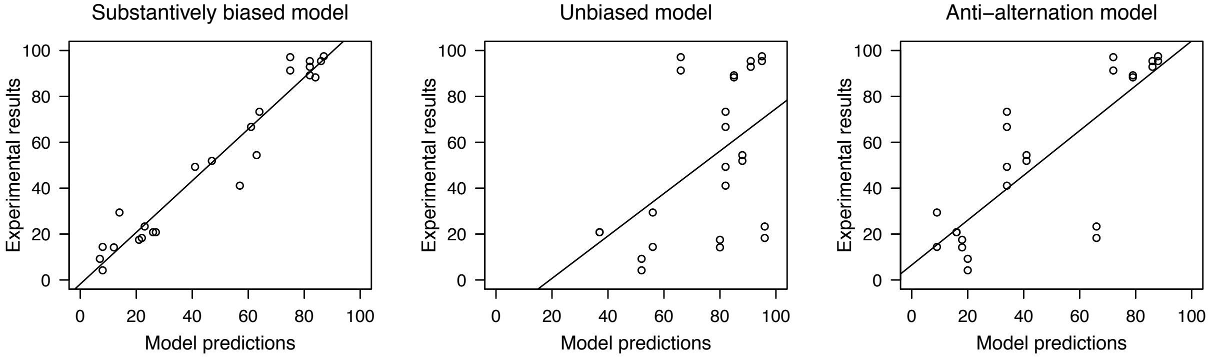

FIGURE 4. Predictions of each model plotted against the experimental results, for each of the observations above. Fitted regression lines are also included.

Model	r^2	Log likelihood
Substantively biased	.94	–1722
Unbiased	.25	–2827
Anti-alternation	.67	–1926

TABLE 5. Proportion of variance explained (r^2) and log likelihood, fitting each model's predictions to the experimental results.

6. Considering other possibilities.

6.1. Effect of different σ^2 values. As implemented, the only free parameter in the model is the squared standard deviation, σ^2, of the prior distribution for each constraint. Recall that the value of σ^2 determines how tightly constraints are bound to their preferred weights (i.e. the μ of the prior distribution). Lower values of σ^2 mean that more data are required to pull the weights away from μ, whereas higher values of σ^2 mean that the weights have more freedom to change in light of the training data.

I had no *a priori* assumptions about how to set σ^2, so several values for σ^2 were tested. To get the predictions reported in §5, σ^2 was set to 0.6, the value that maximized the proportion of variance explained by the substantively biased model (r^2) when fitted to the experimental results. But it is worth considering how different values of σ^2 affect each of the models' performance.

Figure 5 shows the proportion of variance explained (r^2) by each of the three models as a function of different values for σ^2. The most striking aspect of the figure is that for all but the most extreme values of σ^2, the substantively biased model outperforms the anti-alternation model by a considerable margin, and it outperforms the unbiased model by an even greater margin. Thus, the overall conclusion that the substantively biased model outperforms the other models is not dependent on choosing any particular value for σ^2.

Looking at the substantively biased model, we see that the model performs best between the σ^2 values of 0.5 and 0.7, the range at which r^2 reaches a virtual plateau around .94. The reason is that these values of σ^2 represent the 'Goldilocks' range that is 'just right' (at least for this particular learning scenario): the values are low enough that the prior can still have a substantial effect on the outcome but high enough that the training data also have a substantial effect. As the value of σ^2 decreases from 0.5, we see that the model's performance begins to drop, with the drop becoming more abrupt as the value of σ^2 decreases to 0.2 and below. This decrease in performance occurs because as σ^2 drops, the prior becomes too strong such that the training data have little effect on the constraint weights. On the other side, as the value of σ^2 increases from 0.8 and beyond, the model's performance continues to decline at a gradual rate. At an extreme value of 100,000, the prior has very little practical effect on the constraint weights, leaving the weights to be almost entirely dictated by the training data. As a result, all three models converge at similar predictions and thus have very similar r^2 values.

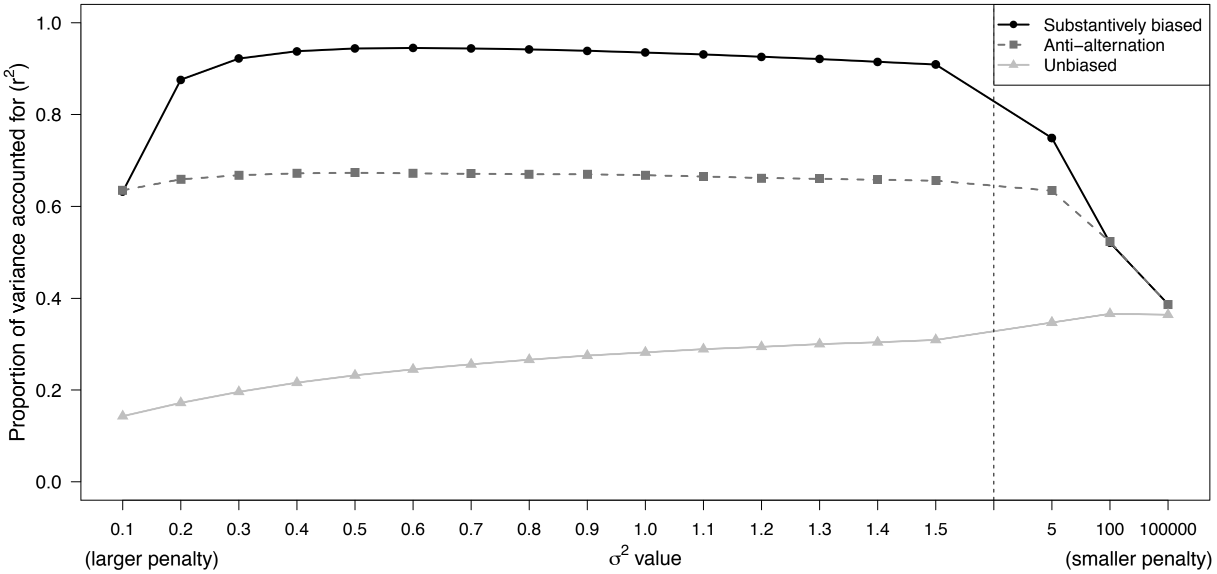

FIGURE 5. Proportion variance explained (r^2) by the substantively biased model, the anti-alternation model, and the unbiased model, according to the value of σ^2.

6.2. EFFECT OF DIFFERENT CONFUSION DATA USED TO DERIVE THE PRIOR. Given that the prior weights are calculated directly from confusion probabilities, the prior weights, and thus the model's performance, will vary depending on which confusion data are used as input. Implicit in this design is the prediction that real language learners will learn slightly differently depending on their own perceptual experience and linguistic background. This strikes me as a reasonable assumption.

For the model predictions reported above, I added together the confusion data listed in Table 2 and Table 3 from Wang and Bilger 1973 (henceforth WB). These data were chosen for several reasons. First, WB's study included all of the crucial consonants tested in White 2014 (i.e. [p, b, f, v, t, d, θ, ð]). Second, WB's participants were native English speakers, like those in White 2014. Third, the consonants were tested in CV syllables (WB:Table 2) and VC syllables (WB:Table 3). The target sounds in White 2014 were located in a VCV context, so I added the data from the CV table and the VC table as a compromise (there was no VCV context collected in WB). Lastly, WB presented the stimuli in noise; the data from WB's Table 2 and Table 3 were summed across all of the signal-to-noise ratios (SNRs) that they tested (i.e. six SNRs ranging from –10 to +15).

Even though the experimental stimuli in White 2014 were presented without noise, I chose to use confusion data from experiments with noise for two reasons. First, the target sounds are all phonemes in English; thus, in clear speech, there are often too few confusions to reliably make assumptions about the relative similarity between sounds. For example, Singh and Black (1966) collected confusion data for English consonants in a VCV context without noise (i.e. the precise conditions in the experiments), but native English listeners made very few errors in that scenario (e.g. there were zero errors when the stimulus was [p] or [b]). Thus such data are not useful for generating the prior weights because they cannot differentiate pairs of sounds according to their similarity.

Second, it is reasonable to assume that experimental participants use their overall experience as listeners throughout their lifetime as the basis of their P-map. In real life, people mostly hear speech in noisy environments. Only rarely do people hear speech that is comparable to clear, carefully pronounced laboratory speech. Thus, using confusion data in noise as the basis of the prior is arguably more appropriate at any rate.

Despite these considerations, the reader may be curious how dependent these results are on using any particular set of confusion data. To address this concern, I ran several versions of the substantively biased model, each with different confusion matrices used as input for the prior. These models are summarized in Table 6. Note that confusion data from WB without noise are also included in this table. In that case, WB tested stimuli of different volumes (ranging from 20 dB to 115 dB) without noise. There were enough errors (due to the stimuli at lower volumes) to differentiate the sounds, so these matrices are included in Table 6.

As expected, model performance varies according to the precise confusion data used as the basis of the prior. Again, it is arguably a good property of the model that its predictions vary depending on its perceptual 'experience'. Crucially, even as different confusion matrices are used to index perceptual similarity, the model's performance remains high. In particular, regardless of which confusion matrix is used, the substantively biased model always outperforms the unbiased and anti-alternation models: the lowest r^2 for the substantively biased model is .77 compared to .25 for the unbiased model and .67 for the anti-alternation model.

In sum, I conclude that although the precise predictions change depending on which confusion data are used to generate the prior, the overall success of the model is not dependent on using any particular confusion matrix.

Source	Table #	Context	In noise?	r^2
WB 1973	2–3[a]	CV and VC	white noise	.94
WB 1973	2	CV	white noise	.93
WB 1973	3	VC	white noise	.92
WB 1973	6–7	CV and VC	none	.93
WB 1973	6	CV	none	.82
WB 1973	7	VC	none	.96
MN 1955	2–6	CV	white noise	.94
C-etal 2004	----[b]	CV and VC	babbled noise	.82
C-etal 2004	----	CV	babbled noise	.79
C-etal 2004	----	VC	babbled noise	.77
Unbiased model (for comparison)				.25
Anti-alternation model (for comparison)				.67

TABLE 6. Performance of models (*r^2*) using different confusion data as the basis for the prior. The shaded lines represent the models reported above: the substantively biased model (top line) and the unbiased and anti-alternation models (bottom lines). In all models, σ^2 is set to 0.6. WB = Wang & Bilger 1973; MN = Miller & Nicely 1955; C-etal = Cutler et al. 2004.

[a] Where multiple table numbers are given, the values were summed across those tables.

[b] The confusion matrices used from Cutler et al. 2004 are not taken from those reported in the paper, which only include the results for one of the SNRs that they tested. Instead, the data were taken from the supplemental webpage for their article, available (June 2013) at http://www.mpi.nl/world/persons/private/anne/materials.html. The matrices available at that webpage include data summed across all SNRs tested.

6.3. FEATURE-BASED CORRESPONDENCE CONSTRAINTS INSTEAD OF *MAP. Thus far, I have assumed *MAP constraints banning correspondence between specific segments, and I have used these constraints as the basis for implementing the substantive bias. However, it is possible to implement the same kind of substantive bias using feature-based correspondence constraints (e.g. traditional IDENT constraints). Under such a scheme, constraints banning changes of more salient features could be assigned higher prior weights than those banning changes of less salient features. Do we need the greater specificity of the *MAP constraints?

To address this question, I replaced the *MAP constraints with two IDENT constraints: IDENT(voice) and IDENT(continuant). These constraints were assigned prior weights by modeling the confusion data from Wang & Bilger 1973, following the same procedure described for the *MAP constraints in §3.3. The resulting prior weights (μ) were 2.05 for IDENT(voice) and 1.28 for IDENT(continuant); these weights reflect the fact that voicing differences were more salient in the Wang & Bilger confusion data overall compared to continuancy differences.

I tested the model on the same input forms from Experiments 1 and 2 used above. Figure 6 shows the model's predictions for all experimental conditions compared to the observed experimental results. Overall, the fit of the model's predictions to the experimental data (r^2 = .62, log likelihood = –2055) was much worse compared to the predictions of the substantively biased model with *MAP constraints (see Table 5).

The reason for the model's poor performance is not surprising. In the Potentially Saltatory condition of Experiment 1 and the Saltatory condition of Experiment 2, the trained alternations (p ⟶ v; t ⟶ ð) lower the weights of both IDENT(voice) and IDENT(continuant) to near 0 while raising the weights of the markedness constraints *V[–voice]V and *V[–cont]V. Because the markedness constraints also motivate changing the intermediate sounds [b, d, f, θ] and the correspondence constraints are too weak to protect these sounds from changing, the model overgeneralizes to untrained sounds (Exp. 1 and 2) and predicts far too many errors in the Saltatory condition of Experiment 2. This problem is the same one that causes classical OT to be unable to generate saltation (as discussed in §2): without expanding the constraint set beyond traditional feature-based IDENT constraints, [p] ⟶ [v] necessarily implies [b] ⟶ [v].

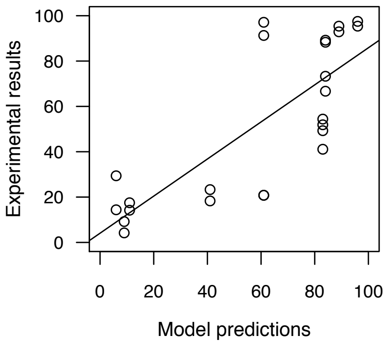

FIGURE 6. Predictions of the substantively biased model with IDENT constraints instead of *MAP constraints, plotted against the experimental results. A fitted regression line is also included.

7. LEARNING A SALTATORY SYSTEM. It is clear from the experimental work cited above (White 2014, White & Sundara 2014) that human learners have a bias against saltatory alternations – they avoid them when faced with ambiguous data and they make errors when forced to learn explicit saltation. We have also seen that a MaxEnt learning model with a bias based on the P-map exhibits similar behavior when presented with the same training data. However, saltations are attested in real languages (Hayes & White 2015), so it must be possible for children to learn such a system successfully. We must therefore ensure that the model can eventually learn a saltatory system, even if such a system is initially dispreferred.

Let us first consider in greater detail how the constraint weights change throughout the learning process. For simplicity, I will consider only the *MAP constraints for labials, but the ones for coronals behave similarly.

The *MAP constraints begin in a P-map-compliant orientation because their prior weights are calculated on the basis of perceptual confusability. For instance, *MAP(p, v) has a hefty weight (3.65) because [p] and [v] are quite dissimilar, *MAP(p, b) has a medium weight (2.44), and *MAP(b, v) has a small weight (1.30) because [b] and [v] are very similar. Recall that in order to have a saltation, these constraints must eventually subvert the P-map-compliant hierarchy such that [p] ⟶ [v] results in a *smaller* penalty relative to [b] ⟶ [v]. Indeed, by taking a look at the weights after the model receives the relatively modest amount of training on explicit saltation from Experiment 2 (i.e. 18 p ⟶ v, 18 t ⟶ ð, 9 b ⟶ b, 9 d ⟶ d), we see that the weights have just reached a non-P-map-compliant orientation (see Table 7; the relevant rows are shaded). Due to the examples of [p] ⟶ [v] in the training data, the weight of *MAP(p, v) begins to plummet (3.65 ⟶ 1.96). At the same time, the initially low weight of *MAP(b, v) is bolstered (1.30 ⟶ 2.02) by the examples of unchanging [b] in the data. Although *MAP(p, v) had a higher weight than *MAP(b, v) in the prior (3.65 vs. 1.30), their weights have now become quite similar (1.96 vs. 2.02) based on evidence from the training data. The weight of *MAP(p, b) is also increased (2.44 ⟶ 2.94), helping to ensure that [p] changes all the way to [v] rather than stopping at intermediate [b].

At the stage of learning based just on the training data from Experiment 2, the grammar is clearly in a transitional state, and this transitional state illustrates why the learner exhibits the anti-saltation learning bias that we see with experimental participants. The training data in Experiment 2, though few in number, are entirely consistent with a saltatory pattern. However, because the model must work against a prior that biases it towards non-saltatory outcomes, the limited amount of training data encountered by the model is not sufficient to completely subvert the constraint hierarchy implicit in the P-map prior. Thus, even though there is no variation in the training data, the ongoing influence of the prior leads the model to predict errors that are comparable to those made by human learners in the experiment.

Constraint	Prior weight	After training data from Experiment 2	After 1000 training data of each type
*V[–voice]V	0	2.45	4.82
*V[–cont]V	0	1.05	0.30
*MAP(p, v)	3.65	1.96	0
*MAP(b, v)	1.30	2.02	4.75
*MAP(p, b)	2.44	2.94	4.62
*MAP(p, f)	1.34	1.74	1.75
*MAP(b, f)	1.96	2.02	2.29
*MAP(f, v)	2.56	2.56	2.56

TABLE 7. Weights for the markedness constraints and the *MAP constraints (labials only) over the course of learning.

Can the model fully overcome the prior? To test whether the model can successfully reach the desired final state (i.e. if it can learn complete, categorical saltation), I trained the model with the following input: 1000 cases of [p] ⟶ [v], 1000 cases of [t] ⟶ [ð], 1000 cases of unchanging [b], and 1000 cases of unchanging [d]. Thus, the type of input was comparable to Experiment 2, but the amount of training data was much more extensive, representing the large amount of input a real child would receive.

The resulting constraint weights are given in the right column of Table 7. The weights of the *MAP constraints continue to move in the same direction. With this much training data, the relevant constraints manage to completely subvert the default hierarchy imposed by the P-map: *MAP(p, v) reaches a weight of 0 (i.e,. it has no effect on the outcome) whereas *MAP(b, v) and *MAP(p, b) reach substantial weights of 4.75 and 4.62, respectively. The large amount of input that goes against the prior leads the model to gradually overcome the prior's influence as it works to successfully account for the observed data.

Ideally, the resulting model should predict [p] ⟶ [v] and [t] ⟶ [ð] nearly 100% of the time, and also that [b] and [d] remain unchanged nearly 100% of the time. The results actually predicted by the model are given in Table 8.

Labials		Coronals	
Outcome	Prediction (in %)	Outcome	Prediction (in %)
p ⟶ v	99.4	t ⟶ ð	99.4
p ⟶ p	0.6	t ⟶ t	0.6
b ⟶ v	1.2	d ⟶ ð	1.1
b ⟶ b	98.8	d ⟶ d	98.9

TABLE 8. Model predictions when provided with 1000 cases of each observation during training, showing that the model can learn a saltatory system.

The model predicts that the saltatory system will indeed be learned successfully, with virtually equal predictions for the labials and coronals. Only 0.6% of the time does the model predict a mistake on the saltatory change, which is low enough to be due to occasional speech errors.[14] Likewise, the model predicts that the intermediate sounds will be changed in error only 1.2% of the time for labials and 1.1% of the time for coronals. Again, this is low enough that

[14] With even more training data, this percentage would get even lower, but never down to 0%. It is not possible for an output to have a prediction that is truly 0% in MaxEnt; it can, however, reach such a low number that the predicted probability is practically 0% (i.e. so low that the output might never occur in a lifetime).

such errors could be considered speech errors. Bolognesi (1998:36) reports that native speakers of Campidanian Sardinian do occasionally spirantize intervocalic voiced stops (in error), but that such errors occur only rarely.

As a final note, there is a plausible connection between learnability and typology: learning biases may serve as a subtle force pushing language change in certain directions over time (e.g. see Moreton 2008, Culbertson et al. 2013). We can hypothesize that the dispreferred status of saltation during learning plays a role in its apparent cross-linguistic rarity and instability (though other factors, such as strength of phonetic precursors, likely play a role as well; e.g. see Moreton 2008, Hayes & White 2015). However, if the bias can be overcome given sufficient input, as we saw in this section, then we might wonder why the bias is not ALWAYS overcome in the real world, given that children learning a language have large amounts of input. This question faces many Bayesian models that implement 'soft' learning biases via a prior. The answer most likely lies in gaining a better understanding of how the model of an individual learner should be integrated within a larger model of language learning and language change within a speech community over time. These issues must be left for future work.

8. COMPARISON WITH WILSON'S (2006) IMPLEMENTATION. The approach to biased phonological learning taken here follows the general approach taken by Wilson (2006): use the prior of the MaxEnt model to implement a substantive bias. However, our approaches differ in the details of the implementation.

Wilson was interested in predicting different rates of velar palatalization (i.e. [k] ⟶ [tʃ] and [g] ⟶ [dʒ]) depending on whether the following vowel was [i], [e], or [a]. Perceptually, [k] and [tʃ] are most similar before [i] and least similar before [a]. In addition, [g] and [dʒ] are less similar than [k] and [tʃ], all else being equal. Typological observations are consistent with the predictions of the P-map: velar palatalization is most common before high vowels and least common before low vowels; it is also more likely to affect [k] than [g].

Wilson implements the substantive bias by setting different σ^2 values for the various markedness constraints he uses to motivate palatalization. In the current paper, by contrast, recall that the substantive bias was implemented by setting a different μ for each faithfulness constraint (i.e. *MAP constraint). For clarity, I will refer to these two implementations as 'Wilson's model' and the 'biased *MAP model', respectively.

Wilson included 12 markedness constraints in his model. These markedness constraints motivate palatalization by penalizing sequences of [k] or [g] followed by a specific vowel (i.e. *ki, *ke, *ka, *gi, *ge, *ga) or a general class of vowels (i.e. *kV$_{[-\text{low}]}$, *kV$_{[-\text{high}]}$, *kV, *gV$_{[-\text{low}]}$, *gV$_{[-\text{high}]}$, *gV). The μ for each of these constraints was set at 0. The σ^2 for each constraint was calculated based on the perceptual similarity between the penalized input consonant and the palatalized output consonant that would result; for example, the σ^2 for *ki was calculated based on the similarity of [k] and [tʃ] before [i].[15] The resulting set of σ^2 values determined how easily the weight for each markedness constraint could be moved from 0. For example, *ki received a relatively high σ^2 whereas *ka received a lower σ^2 because [k] and [tʃ] are more similar before [i] than before [a]. As a result, the weight of *ki could rise more quickly in the face of training data relative to the weight of *ka, which would ultimately result in a greater tendency to palatalize underlying /ki/ compared to /ka/ assuming equal input data. Wilson's model also contained two faithfulness constraints, one penalizing changes to /k/ and one penalizing changes to /g/, which were assigned high (but otherwise fairly arbitrary) values for μ and σ^2.

An important issue for modeling Wilson's experimental results was predicting generalization, for instance that participants who learned palatalization before mid vowels would generalize to the high vowel context but not *vice versa*. Generalization in the model was driven by the set of markedness constraints targeting [k] or [g] in general contexts, such as *kV$_{[-\text{low}]}$. With a Gaussian prior, MaxEnt models prefer to spread responsibility between several constraints rather than putting all of the weight onto a single constraint. For instance, cases of /ke/ ⟶ [tʃe] in training would boost the weight of the general constraint *kV$_{[-\text{low}]}$ in addition to the more specific constraint *ke. As a result, rates of /ki/ ⟶ [tʃi] would be increased at test even if /ki/ never appeared during training.

To see if the biased *MAP approach can also account for Wilson's experimental results, I ran a version of the model equipped with constraints relevant to Wilson's palatalization

[15] Perceptual similarity was calculated by Wilson using the generalized context model of classification (GCM; Nosofsky 1986), taking into account featural similarity, acoustic similarity (peak spectral frequency), confusability (based on confusion data from Guion 1998), and overall response bias. See Wilson 2006 for a detailed account.

experiment. The constraint set and prior weights are shown in Table 9. The model contained two markedness constraints to motivate palatalization, *kV and *gV, which penalize [k] or [g], respectively, when they occur before a vowel. It also contained a set of *MAP constraints banning palatalization in the phonological contexts relevant for Wilson's experiment (i.e. before the vowels [i, e, a]). In total, the model had 8 constraints compared to Wilson's 14 constraints.

Constraint	Prior weight (μ)
*kV	0
*gV	0
*MAP(k, tʃ)/_i	0.21
*MAP(k, tʃ)/_e	0.98
*MAP(k, tʃ)/_a	1.87
*MAP(g, dʒ)/_i	1.22
*MAP(g, dʒ)/_e	1.66
*MAP(g, dʒ)/_a	2.27

TABLE 9. Prior weights (μ) for *MAP constraints based on confusion data in Guion (1998).

To get the prior weights for the *MAP constraints, I used the confusion data from Guion 1998, reported also in Wilson 2006; these are the same confusion data used by Wilson. Guion reports confusions for [k], [tʃ], [g], and [dʒ] before [i] and [a]. Following Wilson, I interpolated to get values for the pre-[e] context; specifically, I took the average number of confusions for the relevant sounds when they occurred before [i] and [a]. Running these through a MaxEnt model, as described in §3.3, resulted in the weights in Table 9, which were entered as the prior μ values for the constraints in the learning model. We can see that the weights reflect the expected similarity relationships: the velar and palatalized sounds are more similar before [i] and least similar before [a], and [k] and [tʃ] are more similar than [g] and [dʒ]. All constraints were assigned a σ^2 of 0.6, the same value used in the models reported in §5.

The overall proportion of variance explained by the biased *MAP model's predictions (r^2) is reported in Table 10 for each of Wilson's four conditions (critical test items). The r^2 values reported by Wilson for his substantively biased model are also provided for comparison. Overall, the predictions of the biased *MAP model represent an excellent fit to Wilson's experimental results; the model actually outperforms Wilson's model in three of the four conditions. The exception is in the Mid condition of Experiment 1, where the biased *MAP model's predictions provide a poor fit to the experimental results. However, as others have pointed out (e.g. Moreton & Pater 2012), Wilson's results in this condition are problematic. Participants who learned palatalization before mid vowels generalized to the high vowel context (as predicted), but also to the low vowel context. Generalization to the low vowel context was unexpected; it is inconsistent with predictions based on typology and the P-map, and participants were even trained that palatalization should not occur before low vowels. It remains unclear why the results in this condition turned out this way, and thus whether the model SHOULD be matching those particular results at all. The biased *MAP model was, however, successful at predicting the HYPOTHESIZED results in the Mid condition.

Condition	r^2 reported for Wilson's model	r^2 for the biased *MAP model
Exp. 1 – High condition	.76	.92
Exp. 1 – Mid condition	.58	.12
Exp. 2 – Voiceless condition	.48	.97
Exp. 2 – Voiced condition	.69	.93

TABLE 10. Proportion of variance accounted for (r^2) by Wilson's (2006) substantively biased model and by the substantively biased *MAP model, when the model predictions are fitted to Wilson's experimental results (critical test items).

In sum, these findings suggest that the approach taken here to implementing the substantive bias, though different than Wilson's approach, is nevertheless successful at capturing the overall observations about velar palatalization presented in Wilson 2006. The biased *MAP model outperforms Wilson's model in three of four conditions (the last of which had curious results), and it does so with fewer constraints. Finally, it is worth noting that Wilson's implementation of the substantive bias as a property of the markedness constraints requires an extension of the traditional role of markedness. In Wilson's model, the markedness constraints must access more than just surface characteristics; they must also have access to the perceptual relationship between the faithful candidate (meaning they must first know which candidate is faithful) and one of the competitor candidates (cf. targeted constraints, Wilson 2001). In contrast, by putting the bias on the faithfulness side, as in the biased *MAP model, this perceptual relationship is assessed by constraints that already evaluate correspondence relationships between two segments.

9. THE INITIAL STATE. The initial state refers to the (presumably innate) state of the child's grammar before any learning occurs. In a MaxEnt model, the prior is often taken to represent the initial state (e.g. see Goldwater & Johnson 2003), so it is worth considering what the model presented here assumes about the initial state of the grammar. Of course, the prior in a MaxEnt model is not merely an 'initial' state. It represents a bias that continues to affect learning throughout the lifetime, as opposed to a default setting that has no lasting effect once learning has commenced (cf. the Gradual Learning Algorithm; Boersma & Hayes 2001).

Previous researchers have argued that in the initial state, it must be the case that markedness outranks faithfulness (Gnanadesikan 1995, Smolensky 1996, Prince & Tesar 1999, Boersma & Levelt 2000, Curtin & Zuraw 2002, Hayes 2004; but cf. Hale & Reiss 1996). As these researchers point out, a major argument for having MARKEDNESS >> FAITHFULNESS in the initial grammar is that children's early, non-adultlike productions appear to reflect principles of markedness, which would be difficult to explain if faithfulness were highly ranked in the grammar.

In the substantively biased instantiation of the model reported above, markedness constraints were set with a μ of 0 whereas the *MAP correspondence constraints were all set with non-zero prior weights. This choice does appear to bear some importance for the model's performance. Raising the μ of the markedness constraints to be higher than the highest *MAP μ causes problems; specifically, the weights of the markedness constraints never have a reason to decrease so the model overgeneralizes.

This dilemma, however, is easily resolved by assuming that the *MAP constraints are evaluated as output-output correspondence constraints (Benua 1997). All of the experiments considered here involved alternations in a paradigm (singular/plural forms of nouns), meaning that paradigm uniformity (Steriade, 2000) is relevant. Thus, the *MAP constraints can be evaluated as output-output constraints (*MAP-OO) rather than input-output constraints (*MAP-IO). Moreover, as mentioned in §2.1, we may have independent reasons for favoring *MAP-OO to *MAP-IO. *MAP constraints must have access to the perceptual similarity of the two forms in correspondence, and it seems conceptually odd to calculate the perceptual similarity between an abstract underlying form and a surface form (see Zuraw 2013).

Several people have claimed that OO-faithfulness constraints are highly ranked in the initial state because there appears to be a natural bias in favor of consistent paradigms. Hayes (1997:46) argues that OO-faithfulness constraints (which he calls Paradigm Uniformity constraints) are undominated in the initial state, appealing to evidence that language change tends to go in the direction of paradigm leveling (see McCarthy 1998 for a similar view). Moreover, Tessier (2006, 2012) and Do (2013) showed that children are biased towards non-alternation, suggesting that OO-faithfulness is highly ranked in their grammars.

The modeling presented here provides further evidence supporting the role of paradigm uniformity as a learning bias. Recall that the anti-alternation model (in which the *MAP constraints all had equal, non-zero weights) outperformed the unbiased model (in which the *MAP constraints, like the markedness constraints, had a prior weight of 0) by a considerable margin. Thus, just having a bias against any alternation at all greatly improved the model's performance. However, the substantively biased model performed even better than the anti-alternation model, suggesting that the substantive P-map bias improved the model over and above just having a bias against alternations more generally.

In conclusion, the account of acquisition proposed here could be summarized as follows. Children begin hearing speech sounds at birth (if not before). After months of experience hearing speech sounds in many environments, they begin to fill in their own P-map with knowledge about the relative similarity of pairs of speech sounds. Some may hold that this knowledge is innate, but that assumption is likely not necessary; though the construction of a P-map is likely universal, the precise contents of the P-map need not be. After building a lexicon during the first year of life, infants begin learning morphology and start learning that the same lexical items can appear in multiple morphophonological contexts – that is, they start learning paradigms. At this point, they already have a natural preference for paradigm uniformity (i.e. *MAP-OO constraints will be preferentially ranked high as they are induced). These *MAP-OO constraints will receive prior weights according to the P-map that has developed from the child's perceptual experience. These weights can then be altered through learning just like the weights of other constraints.

10. CONCLUSION. To summarize, we have seen that saltation is problematic for traditional phonological frameworks. First, saltation is attested in real languages, so it must be possible for children to learn a saltatory system. Any theory that cannot generate saltations, such as classical OT, cannot account for the existence of these patterns. Second, artificial grammar experiments have demonstrated that learners are biased against saltatory patterns when learning phonological alternations, both in terms of how they generalize (White 2014, Exp. 1) and how they learn explicit saltation (White 2014, Exp. 2).

In light of these observations, phonological theory must be able to account for both the existence and the dispreferred status of saltation. Here, I have outlined a framework with the following components, which together allow the model to succeed.

1) *MAP CONSTRAINTS: Saltation is not derivable with only traditional IDENT constraints. The family of *MAP correspondence constraints, adopted from Zuraw (2007, 2013), allows saltation to be learned by making it possible for correspondences between dissimilar sounds to be preferred over correspondences between similar sounds. The *MAP constraints also provide a straightforward way of implementing the P-map bias, which serves to constrain their behavior.

2) P-MAP BIAS: A learning bias based on Steriade's (2009 [2001]) P-map and the principle of minimal modification accounts for the saltation avoidance effect observed in the experiments. In particular, it explains why learners would generalize alternations to phonetically intermediate sounds, but not to other nearby sounds. It also accounts for why learners erroneously change intermediate sounds when learning explicitly saltatory alternations.

3) MAXENT LEARNING**:** The architecture of the MaxEnt learning model is the final crucial component. The prior term serves as an effective vehicle for implementing the P-map bias computationally. Recall that the prior values used to implement the bias were generated directly from experimental confusion data, without being arbitrarily manipulated by hand. From there, the learning process itself is the reason why the model initially exhibits a bias (due to the prior), but is also able reach the final state of (effectively) categorical saltation after sufficient amounts of training data have been observed (§7). The MaxEnt framework allows the prior to be

overturned GRADUALLY through learning, much in the same way that (we can hypothesize) the child would learn saltation.

Looking at the predictions and results from both experiments, the anti-alternation model performs much better than the unbiased model, suggesting that just having non-zero prior weights for the *MAP constraints (i.e. a general bias in favor of paradigm uniformity) provides some benefit. However, the substantively biased model performs much better than the anti-alternation model, indicating that the P-map bias plays a role above and beyond a general bias against alternation. Crucially, only the substantively biased model correctly predicted the basic anti-saltation effect in White's (2014) Experiments 1 and 2. Overall, these findings support the view that substantive bias plays a role in the learning of phonological alternations. The substantively biased model proposed here puts forth a hypothesis about how we should represent this role within phonological theory. The model also provides an implemented framework that can be used in future studies to further explore the role that perceptual similarity plays in phonological learning.

REFERENCES

ANDERSON, STEPHEN R. 1981. Why phonology isn't "natural". *Linguistic Inquiry* 12.493–547.

ANTTILA, ARTO. 1997. *Variation in Finnish phonology and morphology*. Stanford, CA: Stanford University dissertation.

ARCHANGELI, DIANA, and DOUGLAS PULLEYBLANK. 1994. *Grounded Phonology*. Cambridge, MA: MIT Press.

BAER-HENNEY, DINAH, and RUBEN VAN DE VIJVER. 2012. On the role of substance, locality, and amount of exposure in the acquisition of morphophonemic alternations. *Laboratory Phonology* 3.221–50.

BAKOVIĆ, ERIC. 2004. Unbounded stress and factorial typology. *Optimality Theory in phonology: A reader*, ed. by John McCarthy, 202–14. London: Blackwell.

BECKER, MICHAEL; NIHAN KETREZ; and ANDREW NEVINS. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87.84–125.

BECKER, MICHAEL; ANDREW NEVINS; and JONATHAN LEVINE. 2012. Asymmetries in generalizing alternations to and from initial syllables. *Language* 88.231–68.

BERGER, ADAM L.; VINCENT J. DELLA PIETRA; and STEPHEN A. DELLA PIETRA. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22.39–71.

BENUA, LAURA. 1997. *Transderivational identity: Phonological relations between words.* Amherst: University of Massachusetts dissertation.

BLEVINS, JULIETTE. 2004. *Evolutionary phonology: The emergence of sound patterns.* Cambridge: Cambridge University Press.

BOERSMA, PAUL, and BRUCE HAYES. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32.45–86.

BOERSMA, PAUL, and CLARA LEVELT. 2000. Gradual constraint-ranking learning algorithm predicts acquisition order. *Proceedings of Child Language Research Forum* 30.229–37.

BOERSMA, PAUL, and JOE PATER. 2008. Convergence properties of a gradual learning algorithm for Harmonic Grammar. Amherst: University of Massachusetts, MS. [ROA 970]. http://roa.rutgers.edu/article/view/1000.

BOLOGNESI, ROBERTO. 1998. *The phonology of Campidanian Sardinian: a unitary account of a self-organizing structure*. The Hague: Holland Institute of Generative Linguistics.

CALAMARO, SHIRA, and GAJA JAROSZ. 2015. Learning general phonological rules from distributional information: A computational model. *Cognitive Science* 39.647–66.

CARPENTER, ANGELA C. 2010. A naturalness bias in learning stress. *Phonology* 27.345–92.

CHAMBERS, KYLE E.; KRISTINE H. ONISHI; and CYNTHIA FISHER. 2011. Representations for phonotactic learning in infancy. *Language Learning and Development* 7.287–308.

CRISTIA, ALEJANDRINA; JEFF MIELKE; ROBERT DALAND; and SHARON PEPERKAMP. 2013. Constrained generalization of implicitly learned sound patterns. *Journal of Laboratory Phonology* 4.259–85.

CRISTIÀ, ALEJANDRINA, and AMANDA SEIDL. 2008. Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development* 4.203–27.

CULBERTSON, JENNIFER; PAUL SMOLENSKY; and COLIN WILSON. 2013. Cognitive biases, linguistic universal, and constraint-based grammar learning. *Topics in Cognitive Science* 5.392–424.

Curtin, Suzanne, and Kie Zuraw. 2002. Explaining constraint demotion in a developing system. *Proceedings of the 26th Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press.

Cutler, Anne; Andrea Weber; Roel Smits; and Nicole Cooper. 2004. Patterns of English phoneme confusion by native and non-native listeners. *Journal of the Acoustical Society of America* 116.3668–78.

Do, Youngah. 2013. *Biased learning of phonological alternations.* Boston: MIT dissertation.

Eisner, Jason. 2000. Review of Kager: “Optimality Theory”. *Computational Linguistics* 26.286–90.

Finley, Sara. 2008. *The formal and cognitive restrictions on vowel harmony.* Baltimore: Johns Hopkins University dissertation.

Finley, Sara, and William Badecker. 2012. Learning biases for vowel height harmony. *Journal of Cognitive Science* 13.287–327.

Fleischhacker, Heidi. 2005. *Similarity in phonology: Evidence from reduplication and loan adaptation*. Los Angeles: UCLA dissertation.

Gnanadesikan, Amalia E. 1995. Markedness and faithfulness constraints in child phonology. Amherst: University of Massachusetts, ms. [ROA 67]. http://roa.rutgers.edu/article/view/68.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation within Optimality Theory,* ed. by Jennifer Spenader; Anders Eriksson, and Östen Dahl, 111–20. Stockholm: Stockholm University Department of Linguistics.

Gordon, Matthew. 2002. A factorial typology of quantity insensitive stress. *Natural Language and Linguistic Theory* 20.491–552.

GUION, SUSAN G. 1998. The role of perception in the sound change of velar palatalization. *Phonetica* 55.18–52.

HALE, MARK, and CHARLES REISS. 1996. The initial ranking of faithfulness constraints in UG. Montreal: Concordia University, MS. [ROA 104]. http://roa.rutgers.edu/article/view/115.

HALE, MARK, and CHARLES REISS. 2000. ‘Substance abuse’ and ‘dysfunctionalism’: Current trends in phonology. *Linguistic Inquiry* 31.157–69.

HAYES, BRUCE. 1997. Anticorrespondence in Yidiny. Los Angeles: UCLA, MS.

HAYES, BRUCE. 1999. Phonetically-driven phonology: The role of Optimality Theory and Inductive Grounding. *Functionalism and formalism in linguistics,* vol. 1, ed. by Michael Darnell, Edith Moravscik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatly, 243–85. Amsterdam: John Benjamins.

HAYES, BRUCE. 2004. Phonological acquisition in Optimality Theory: the early stages. *Fixing Priorities: Constraints in Phonological Acquisition*, ed. by Rene Kager, Joe Pater, and Wim Zonneveld, 158–203. Cambridge: Cambridge University Press.

HAYES, BRUCE, and MARGARET MACEACHERN. 1998. Quatrain form in English folk verse. *Language* 64.473–507.

HAYES, BRUCE, and DONCA STERIADE. 2004. Introduction: The phonetic basis of phonological markedness. *Phonetically-based phonology*, ed. by Bruce Hayes, Robert Kirchner, and Donca Steriade, 1–32. Cambridge: Cambridge University Press.

HAYES, BRUCE, and JAMES WHITE. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44.45–75.

HAYES, BRUCE, and JAMES WHITE. 2015. Saltation and the P-map. *Phonology* 32.1–36.

HAYES, BRUCE, and COLIN WILSON. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440.

HAYES, BRUCE; COLIN WILSON; and ANNE SHISKO. 2012. Maxent grammars for the metrics of Shakespeare and Milton. *Language* 88.691–731.

HAYES, BRUCE; KIE ZURAW; PÉTER SIPTÁR; and ZSUZSA LONDE. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85.822–63.

HELLBERG, STAFFAN. 1978. Unnatural phonology. *Journal of Linguistics* 14.157–77.

ITO, JUNKO, and ARMIN MESTER. 2003. On the sources of opacity in OT: coda processes in German. *The Syllable in Optimality Theory*, ed. by Caroline Féry and Ruben van de Vijver, 271–303. Cambridge: Cambridge University Press.

JOHNSON, MARK. 2002. Optimality-theoretic lexical functional grammar. *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*, ed. by Paola Merlo and Suzanne Stevenson, 59–74. Amsterdam: John Benjamins.

KAUN, ABIGAIL. 1995. *The typology of rounding harmony: An optimality theoretic approach.* Los Angeles: UCLA dissertation.

KAWAHARA, SHIGETO. 2006. A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese. *Language* 82.536–74.

KIPARSKY, PAUL. 1982. *Explanation in Phonology.* Dordrecht: Foris.

LEGENDRE, GÉRALDINE; YOSHIRO MIYATA; and PAUL SMOLENSKY. 1990. Harmonic Grammar — A formal multi level connectionist theory of linguistic well formedness: Theoretical foundations. *Proceedings of the 12th Annual Conference of the Cognitive Science Society.* 388-95.

LUBOWICZ, ANNA. 2002. Derived environment effects in Optimality Theory. *Lingua* 112.243–80.

MAGRI, GEORGIO. 2012. Convergence of error-driven ranking algorithms. *Phonology* 29.213–69.

MARTIN, ANDREW. 2011. Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87.751–70.

MCCARTHY, JOHN. 1998. Morpheme structure constraints and paradigm occultation. *Proceedings of the Chicago Linguistic Society* 35.123–50). Chicago, CLS.

MCCARTHY, JOHN. 2003. Comparative markedness. *Theoretical Linguistics* 29.1–51.

MCCARTHY, JOHN, and ALAN PRINCE. 1995. Faithfulness and reduplicative identity. *Papers in optimality theory* (University of Massachusetts occasional papers in linguistics 18), ed. by Jill Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk, 249–384. Amherst: University of Massachusetts Deparment of Linguistics.

MIELKE, JEFF. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press.

MIELKE, JEFF. 2012. A phonetically based metric of sound similarity. *Lingua* 122.145–63.

MILLER, GEORGE A., and PATRICIA E. NICELY. 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27.338–52.

MORETON, ELLIOTT. 2008. Analytic bias and phonological typology. *Phonology* 25.83–127.

MORETON, ELLIOTT, and JOE PATER. 2012. Structure and substance in artificial-phonology learning. Part II: Substance. *Language and Linguistics Compass* 6.702–18.

NAGY, NAOMI, and BILL REYNOLDS. 1997. Optimality Theory and variable word-final deletion in Faetar. *Language variation and change* 9.37–55.

NEVINS, ANDREW. 2010. Two case studies in phonological universals: A view from artificial grammars. *Biolinguistics* 4.218–33.

NOSOFSKY, ROBERT M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115.39–57.

OHALA, JOHN. 1981. The listener as a source of sound change. *Chicago Linguistic Society* 17.178–203.

OHALA, JOHN. 1993. Sound change as nature's speech perception experiment. *Speech Communication* 13.155–61.

PATER, JOE. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39.334–45.

PATER, JOE. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33.999–1035.

PATER, JOE; ROBERT STAUBS; KAREN JESNEY; and BRIAN SMITH. 2012. Learning probabilities over underlying representations. *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, 62–71. Montreal: Association for Computational Linguistics.

PEPERKAMP, SHARON, and EMMANUEL DUPOUX. 2007. Learning the mapping from surface to underlying representations in an artificial language. *Laboratory Phonology* 9.315–38.

PEPERKAMP, SHARON; ROZENN LE CALVEZ; JEAN-PIERRE NADAL; and EMMANUEL DUPOUX. 2006a. The acquisition of phonological rules: Statistical learning with linguistic constraints. *Cognition* 101.B31–41.

PEPERKAMP, SHARON; KATRIN SKORUPPA; and EMMANUEL DUPOUX. 2006b. The role of phonetic naturalness in phonological rule acquisition. *Proceedings of the Boston University Conference on Language Development* 30.464–75.

PIERREHUMBERT, JANET. 2006. The statistical basis of an unnatural alternation. *Laboratory Phonology VIII: Varieties of phonological competence*, ed. by Louis Goldstein, D. H Whalen, and Catherine Best, 81–106. Berlin: Mouton de Gruyter.

PRESS, WILLIAM H.; BRIAN P. FLANNERY; SAUL A. TEUKOLSKY; and WILLIAM T. VETTERLING. 1992. *Numerical recipes in C: The art of scientific computing.* Cambridge: Cambridge University Press.

PRINCE, ALAN, and PAUL SMOLENSKY. 2004 [1993]. *Optimality theory: Constraint interaction in generative grammar.* Oxford: Blackwell.

PRINCE, ALAN, and BRUCE TESAR. 1999. Learning phonotactic distributions. New Brunswick, NJ: Rutgers University, MS. [ROA 353]. http://roa.rutgers.edu/article/view/363.

PYCHA, ANNE; PAWEL NOWAK; EURIE SHIN; and RYAN SHOSTED. 2003. Phonological rule-learning and its implications for a theory of vowel harmony. *West Coast Conference on Formal Linguistics* 22.423–35.

ROSS, KIE. 1996. *Floating phonotactics: Variability in infixation and reduplication of Tagalog loanwords*. Los Angeles, UCLA M.A. thesis.

SEIDL, AMANDA, and EUGENE BUCKLEY. 2005. On the learning of arbitrary phonological rules. *Language Learning and Development* 1.289–316.

SINGH, SADANAND, and JOHN W. BLACK. 1966. Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. *Journal of the Acoustical Society of America* 39.372–87.

SKORUPPA, KATRIN; ANNA LAMBRECHTS; and SHARON PEPERKAMP. 2011. The role of phonetic distance in the acquisition of phonological alternations. *Proceedings of the 39th Annual Meeting of the North East Linguistic Society*. 717–29.

SKORUPPA, KATRIN, and SHARON PEPERKAMP. 2011. Adaptation to novel accents: Feature-based learning of context-sensitive phonological regularities. *Cognitive Science* 35.348–66.

SMOLENSKY, PAUL. 1996. The initial state and "richness of the base" in Optimality Theory. Baltimore: Johns Hopkins University, MS. [ROA 154]. http://roa.rutgers.edu/article/view/165.

SMOLENSKY, PAUL. 2006. Optimality in phonology II: Harmonic completeness, local constraint conjunction, and feature domain markedness. *The Harmonic Mind: From Neural to Optimality-Theoretic Grammar*, ed. by Paul Smolensky and Géraldine Legendre. Cambridge, MA: MIT Press.

SMOLENSKY, PAUL, and GÉRALDINE LEGENDRE. 2006. *The Harmonic Mind: From Neural Computation to Optimality Theoretic Grammar.* Cambridge, MA: MIT Press.

STERIADE, DONCA. 2000. Paradigm uniformity and the phonetics-phonology boundary. *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, ed. by Michael. B. Broe and Janet. B. Pierrehumbert, 313–34. Cambridge: Cambridge University Press.

STERIADE, DONCA. 2009 [2001]. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. *The Nature of the Word: Studies in Honor of Paul Kiparsky*, ed. by Kristin Hanson and Sharon Inkelas, 151–80. Cambridge: MIT Press.

SUNDARA, MEGHA; YUN JUNG KIM; JAMES WHITE; and ADAM CHONG. 2013. There is not pat in patting: Acquisition of phonological alternations by English-learning 12-month-olds. Talk given at the 38th Boston University Conference on Language Development. (http://www.ucl.ac.uk/~ucjtcwh/index_files/SundaraEtAl_BUCLD2013.pdf)

TESSIER, ANNE-MICHELLE. 2006. Testing for OO-Faithfulness in artificial phonological acquisition. *Proceedings of the 30th Annual Boston University Conference on Language Development*, 619–30. Somerville, MA: Cascadilla Press.

TESSIER, ANNE-MICHELLE. 2012. Testing for OO-Faithfulness in the acquisition of consonant clusters. *Language Acquisition* 19.144–73.

WANG, MARILYN D., and ROBERT C. BILGER. 1973. Consonant confusions in noise: a study of perceptual features. *Journal of the Acoustical Society of America* 54.1248–66.

WHITE, JAMES. 2014. Evidence for a learning bias against saltatory phonological alternations. *Cognition* 130.96–115.

WHITE, JAMES, and MEGHA SUNDARA. 2014. Biased generalization of newly learned phonological alternations by 12-month-old infants. *Cognition* 133.85–90.

WILSON, COLIN. 2001. Consonant cluster neutralization and targeted constraints. *Phonology* 18.147–97.

WILSON, COLIN. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30.945–82.

YU, ALAN C. L. 2004. Explaining final obstruent voicing in Lezgian: Phonetics and history. *Language* 80.73–97.

ZURAW, KIE. 2007. The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. *Language* 83.277–316.

ZURAW, KIE. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory* 28.417–72.

ZURAW, KIE. 2013. *MAP constraints. Los Angeles: UCLA, MS.