Perceptual saliency, lenition, and learnability: An artificial grammar learning study

Bethany Sturman
MA Thesis, UCLA
6/15/18

This paper explores two theoretical frameworks describing intervocalic lenition: a hearer-focused, perception-based framework versus a production-based framework. The perception-based framework (Kingston, 2008; Katz, 2016) posits that intervocalic lenition is primarily driven by a preference for higher intensity within prosodic constituents to aid the listener in distinguishing prosodic phrase boundaries. The production-based framework is an effort-driven theory (Kirchner, 1998, 2004) proposing that speakers wish to minimize articulatory effort when speaking. Articulatory gestures requiring movement that is further and/or faster are dispreferred.

These two frameworks make different predictions about when lenition should occur. The production-based framework predicts that intervocalic lenition should be sensitive to the height of the surrounding vowels, whereas the perception-based framework does not. To distinguish these frameworks empirically, I propose a Poverty of Stimulus artificial grammar learning experiment to explore the extent to which surrounding vowel height influences how readily learners generalize the lenition pattern of spirantization.

The results turn out to support a third hypothesis, perceptual saliency (Steriade, 2001a, 2001b; Fleischhacker, 2005). Learners generalize asymmetrically according to the condition they are trained on. The learner is more likely to generalize the pattern if she learns the alternation in a more perceptually salient position and is asked to generalize to a less salient position than vice versa.

1. Introduction

Lenition is an umbrella term phonologists use to refer to several "weakening" phenomena found cross-linguistically (see Trask, 2004 for an attempt at a definition; Lavoie, 2001for a theoretical survey; and Gurevich, 2005 for a typological survey). Various subtypes of lenition include:

- Voicing: voiceless obstruents become voiced (t → d)
- Debuccalization: oral obstruents become glottal (t → ʔ)
- Degemination: long consonants become short (tː → t)
- Spirantization: stops become continuants (t → θ)
- Tapping: stops become taps (t → ɾ)

Lenition occurs in a number of different environments, most notably in intervocalic position and word final position. Lumping both of these positions under the heading "lenition," however, may actually be collapsing two distinct classes of phenomena into one. This possible misclassification could be complicating and conflating what a tenable theory of lenition must account for. As such, I will focus on intervocalic lenition, the phenomenon Katz terms "continuity lenition"

(Katz, 2016). It is also worth noting that the term "lenition" originated in the study of historical linguistics to refer to diachronic weakening processes observed in language change (Sievers, 1881; Jesperson, 1904; Bloomfield, 1932; Lass and Anderson 1975; Hyman, 1975). This paper, however, focuses exclusively on synchronic alternations.

One question of interest from the perspective of synchronic phonology is what phonological or phonetic forces drive intervocalic lenition. The two prominent theories claim that lenition is either perception-based (Katz 2016, Kingston 2008) or production-based (Kirchner 2004). These frameworks make different empirical predictions about the degree of influence the height of surrounding vowels should have in driving lenition in an intervocalic frame. This paper explores the question of whether experimental data from a learnability study can provide empirical evidence that might aid in distinguishing between these two frameworks. As it turns out, the experimental results point to a third alternative, a perceptual account relating to saliency.

2. The hearer-focused perception-based framework (Katz 2016, Kingston 2008)

In the hearer-focused perception-based framework of lenition, continuity lenition and initial fortition are two sides of the same coin (Keating 2006, Kingston 2008). Specifically, both of the observed phenomena assist listeners in distinguishing prosodic constituents. Prosodic constituents should be of low intensity at the boundaries and high intensity internally. The phonetic phenomenon of initial fortition has been well-documented in the literature. Measures of articulatory closure strength through palatography experiments indicate that the degree of fortition increases according to the strength of the prosodic boundary, with stronger closures corresponding to larger boundaries (Fougeron and Keating, 1997; Keating et al. 2004). While initial fortition decreases intensity at the boundary of the constituent, lenition aims to preserve auditory continuity within prosodic constituents. Specifically, lenition is the result of implementing a minimum level of phrase-internal intensity. Note that intensity and sonority are two distinct scales and should not be confused. The intensity continuum of segments is pictured below:

Less intense ⟵⟶ More intense

voiceless stop	voiceless continuant	voiced stop	voiced continuant	tap	glide

[Katz, 2016]

Unlike sonority, intensity makes several gradable distinctions between various types of obstruents. Moreover, under this model of intensity the lower band of the frequency spectrum is privileged. Voicing makes a larger contribution to the intensity of a segment than sibilant frication. On this scale, a voiced stop [d] is more intense than a voiceless fricative [s].

In Katz's model, the relative intensities of segments combine with their respective durations to create a set of constraints that Katz terms the BOUNDARY-DISRUPTION constraint family. These constraints unify the phenomena of initial fortition and intervocalic lenition. The BOUNDARY-DISRUPTION constraints accomplish this by instituting an intensity threshold. The threshold constitutes a maximum allowable level of intensity at the edges of a prosodic constituent and a minimum allowable intensity phrase-internally.

Due to the complementary relation of the intensity restrictions at the edge of the constituent and phrase-internally, Katz claims that intervocalic lenition is generally a non-neutralizing phenomenon. This prediction is supported by the typology (Gurevich 2005), in which the extremely rare cases of intervocalic lenition neutralizing a contrast are also predicted by Katz's account. One of the most familiar cases of neutralizing intervocalic lenition is English tapping, which neutralizes the voicing contrast available in the alveolar stops. In Katz's analysis, the intensity threshold is set too high on the intensity scale to allow any possible voicing contrast at this place. The non-neutralizing nature of intervocalic lenition also aligns with the phonetic understanding that intervocalic position is the best position for consonant place contrasts due to the presence of all possible acoustic cues (Steriade 2001b).

3. The production-based framework: Effort (Kirchner 2004)

The most obvious alternative to a perception-based framework is a production-based one. Under Kirchner's analysis, lenition is the result of a production-based framework that seeks to reduce articulatory effort. People are generally efficient and want to expend as little effort as possible while speaking. Lenition phenomena are therefore driven by grammatical constraints referring to physical effort. Kirchner (1998) uses a spring-mass model of the tongue and lips to estimate the amount of effort required to make certain gestures. Effort corresponds to the force needed from the muscles to move the articulators to their target or to keep them in position for a gesture. The greater the force required, the more adenosine-triphosphate (ATP) must be used to power the articulators. Gestures requiring more physical effort, which corresponds to having to travel either further or faster, or having to maintain a closure for a longer duration, are dispreferred.

Kirchner models this restriction with the markedness constraint family LAZY(*X*), which penalizes candidates with articulatory gestures more effortful than some articulatory effort threshold *x* (Kirchner 2004). LAZY constraints interact with faithfulness constraints seeking to preserve perceptual distinctions, such as *MAP constraints (Zuraw 2007, 2013). A sample toy analysis of a basic spirantization pattern is shown in the tableaux below.

/ta/	LAZY(t_1)	*MAP(t, θ)	LAZY(t_2)
☞[ta]			*
[θa]		*!	

/ata/	LAZY(t_1)	*MAP(t, θ)	LAZY(t_2)
[ata]	*!		*
☞[aθa]		*	*

In the first tableau, the effort needed to produce [ta] falls between the thresholds t_1 and t_2, resulting in a violation of only the LAZY constraint ranked below *MAP(t, θ), so the faithful stop surfaces. In the second tableau, the effort needed to produce [ata] now exceeds both thresholds t_1 and t_2, and so the higher ranked LAZY constraint drives lenition.

Kirchner suggests that as the field's understanding of articulators deepens, we will have a better understanding of the various muscle groups involved in performing particular articulatory gestures. From this knowledge, the articulatory effort thresholds could be empirically determined, enabling the theory to be implemented on a more formal, quantitative level.

Relating to intervocalic lenition, the effort-based model predicts that classic realizations of lenition, particularly voicing and spirantization, are the result of effort reduction. It takes more effort to stop glottal vibrations and restart them (hence voiceless sounds become voiced intervocalically), and the distance traveled to form a stop closure is greater than to create a fricative (hence spirantization). Because effort in the production-based model is a function of distance traveled and the speed needed to produce a gesture, Kirchner's model predicts that the height of the surrounding vowels in an intervocalic frame should influence the rate of lenition. Specifically, moving from a low back vowel to an alveolar stop requires more effort than moving from a high front vowel to an alveolar stop, and so the former should be dispreferred. This would result in higher rates of lenition of stops between low vowels than lenition of the same stops between high vowels. Kirchner (ch. 6, 1998) refers to this asymmetry in his Aperture Conditioning Generalization. The Aperture Conditioning Generalization can be summarized as follows: when segments with less open apertures drive some sort of lenition in a consonant C, segments with more open apertures must drive at least that level of lenition in C.

Kirchner (1998) identifies a few cases from the typology that seem to lend support to the manifestation of the Aperture Conditioning Generalization (though c.f. Kingston, 2008 for alternative explanations of the data). In Mbabaram (Australian, Dixon 1991), voicing lenition in intervocalic position is more likely to occur following a low vowel (/ a_V) than following a high vowel (/ i _V), which is still more frequent than voicing following [l] (/ l_V). In Chitwan Tharu (Leal 1972), /b/ spirantizes to [β] between non-high vowels. In Sotho (Doke 1957, Grammont 1939), stops spirantize between non-high vowels and /d/ becomes [l] before a non-high vowel. Finally, in Korean (Martin 1992), /w/ is deleted before non-high vowels.

4. A third alternative: Perceptual saliency (Steriade 2001, Fleischhacker 2005)

While not initially considered when designing the lenition learnability experiment, there is a third logical alternative to the Continuity hypothesis and the Effort hypothesis in the possible experimental outcomes. I call this alternative the Saliency hypothesis. This hypothesis was first put forward in the work of Steriade and Fleischhacker. They claim that phonological processes are shaped by pressure to maintain perceptual similarity between corresponding forms (Steriade 2001a, Steriade 2001b, Fleischhacker 2005). Under the Saliency hypothesis, the strength of the faithfulness constraint preventing a sound change corresponds to the perceptual salience of the change. The more salient the change, the stronger the corresponding faithfulness constraint is. Moreover, if a language permits a sound change in a more perceptible position, it should also permit that change in a less perceptible position.

Steriade looks to the Saliency hypothesis as the answer to the Too Many Solutions problem in Optimality Theory (Prince & Smolensky, 2008). Given the richness of the base, there are many possible repairs for any particular markedness violation. Across the typology, however, very few of these repairs are actually implemented. For example, in languages that prohibits word-final voiced obstruents, the repair that is universally employed is final obstruent devoicing (e.g. German, Russian, among many others) rather than nasalization or word-final vowel epenthesis. According to Steriade (2001), languages prefer to use obstruent devoicing because it is the least perceptually salient change available to repair the markedness violation. Similarly, Fleischhacker (2005) appeals to the Saliency hypothesis to explain the outcomes of complex onset simplification in reduplication as well as loanword adaptation. She claims that obstruent-sonorant clusters are more likely to permit an intervening vowel than sibilant-obstruent clusters because the insertion into the obstruent-sonorant cluster is less perceptually salient than in the case of the sibilant-obstruent cluster.

While the Continuity hypothesis is also rooted in perception, it is more focused on perception from the hearer's perspective (i.e. the perception of word boundaries). The Saliency hypothesis, however, is linked to perception more broadly (i.e. how salient a change is permissible before the violation of faithfulness to similar corresponding forms becomes too extreme.) Although this hypothesis is agnostic regarding the particular mechanism (or stochastic process) that drives lenition, it does predict that when lenition patterns emerge, they will begin in environments where the change is less perceptually salient.

5. Experiment motivation

The two frameworks put forth as the phonological impetus for lenition make different empirical predictions about when lenition should occur which can be tested through experimentation. The production-based framework hypothesizes that there is a UG bias disfavoring articulatory effort. As such, the Effort hypothesis predicts that lenition should happen more frequently when the tongue body has a greater distance to travel (i.e. aCa should lenite more frequently than iCi) due to the increased effort involved in producing the stop closure.

The hearer-focused perception-based framework predicts that prosodic constituents should contain lower-intensity boundary segments and higher-intensity medial segments. Accordingly, the Continuity hypothesis predicts that the height of the flanking vowels in a VCV frame should not be a factor in determining the frequency of lenition (i.e. aCa should lenite with the same frequency as iCi).

We can test for a possible UG bias for the production-based framework using a Poverty of Stimulus artificial grammar learning experiment modeled closely after Wilson's (2006). Specifically, if learners are trained that a language has a rule that /iti/ becomes [iθi], they should generalize that rule to the unfamiliar environment [a_a] since spirantizing the stop between low vowels reduces effort more than spirantizing between high vowels. On the other hand, if learners are trained that /ata/ becomes [aθa], the grammar may be such that learners do not generalize to [iti] words since the smaller reduction in effort is outranked by an intervening faithfulness constraint.

The Effort hypothesis predicts that generalization will be asymmetric: learning spirantization between high vowels should generalize to spirantization between low vowels but not necessarily the inverse.

The Continuity hypothesis predicts that the rate of generalization will be the same regardless of whether the rule is learned in the context of high vowels or low vowels. Both vowels will surpass a Boundary-Disruption threshold that drives the spirantization of intervocalic stops, and so the Continuity hypothesis predicts full generalization in both conditions.

The Saliency hypothesis predicts that learners will generalize the rule from a more perceptually salient environment to the less perceptually salient environment but not the inverse. In the case of learning spirantization in the environment of either low or high vowels, generalization will be asymmetric but in the opposite direction of the Effort hypothesis. Specifically, the change from /ata/ to [aθa] is more perceptually salient than from /iti/ to [iθi], and so generalizing to spirantizing between low vowels after being trained on the change between high vowels should be dispreferred. To my knowledge, there is not yet any phonetic perceptibility data available that speaks to the relative perceptibility of these changes directly. This claim seems reasonable, however, in light of the physiology of the vocal tract as well as support from the typology. When producing a high vowel, the tongue forms a narrow channel due to the tongue blade's position very close to the hard palate. This leads to a noisier stop release burst compared to the burst before a low vowel (Kirchner 1998). This noisy stop release burst can lead to a shift towards more continuant-like obstruents. Such a shift is indeed manifested in the typology. Kirchner (1998) notes a few examples of assibilatory affrication specifically before high vowels, such as Québecois French (/t, d/ → [ts, dz] / __ {i, y, ɪ, ʏ}) and Japanese (/t, d/ → [ts, dz] / __ ɨ).

A summary of the predictions of the three hypotheses is shown in Table 1 below:

	Effort	Continuity	Saliency
Learn /ata/ → [aθa]	× (no generalization)	✓	✓?
Learn /iti/ → [iθi]	✓	✓	×?

Table 1: Summary of hypothesis predictions. (✓) indicates generalization is predicted, (×) indicates generalization is not predicted

6. Artificial Grammar Learning: Methodology

6.1 The artificial language

The language consists of nouns for animals which appeared in three forms: singular, plural, and diminutive. The lenition pattern was intervocalic spirantization. The language included vowels [i o a], sonorants [m n r l j] and stops [t d] in complementary distribution with corresponding fricatives [θ ð], so as to reflect the non-neutralizing typology of intervocalic lenition. The language was constructed to be as natural as possible for native English speakers to learn. As such, the obstruent inventory consisted only of coronals. Dorsals were excluded due to the lack of velar fricatives in English. Labials were excluded due to the difference in place of articulation of the stop (bilabial) versus the fricative (labiodental) in English. Moreover, jaw height evidence indicates that [f] and [v] might actually be more effortful than [p] and [b] (Keating et al. 1994). Since the production hypothesis crucially predicts lenition as a result of effort reduction, spirantizing to a more effortful fricative is not a relevant pattern to test the Effort hypothesis.

Stems were disyllabic CVCVC, constructed using the following template. The initial consonant was selected from the consonant inventory (excluding fricatives), the first vowel was chosen at random, and the medial consonant was a sonorant, in order to preserve poverty of stimulus. For target items, the second vowel was either [a] or [i] to construct the left segment of the intervocalic frame of interest, and the final consonant was [t] or [d]. In filler items, the second vowel was also chosen randomly, and the final consonant was a non-glide sonorant. The right half of the intervocalic lenition frame was introduced via the two suffixes, [-al] (plural) and [-in] (diminutive). In the inflected forms, the stem final stops became fricatives and the sonorant final stems were invariant. The stress pattern was penultimate. This stress pattern discourages vowel reduction in the intervocalic lenition frame and matches the typical stress pattern of English nouns (Liberman & Prince 1977).

Following these parameters, items were constructed using a MATLAB script written to balance the distribution of sounds and to ensure the uniqueness of every item. Examples can be seen in Table 2 below. The full list of stimuli is listed in Appendix A.

Training condition	Singular	Plural	Diminutive	Gloss
Both	milom	milomal	milomin	camel
Train on [i]	tonit	toniθal	toniθin	bat
Train on [i]	jolid	joliðal	joliðin	bison
Train on [a]	ralad	ralaðal	ralaðin	elephant
Train on [a]	dolat	dolaθal	dolaθin	cow

Table 2: Example training stimuli. Greyed cells never appear in training phases in order to preserve Poverty of Stimulus

6.2 Task Design

The experiment was run at the UCLA Phonetics Lab in a sound-attenuated room. The task was administered by a phonetically trained experimenter (either the author or an undergraduate research assistant blind to the purposes of the experiment). The in-person method was employed after several iterations of a computerized version of the task. Learning the alternation in the computerized versions proved to be exceedingly difficult for participants. Details of the various methodological iterations can be found in Appendix B.

The experiment was a between-subjects design with two conditions: training on stems with a final [a] vowel (e.g. [dolat]) and generalizing to stems with final [i] (e.g. [nilid]), and vice versa. The task consisted of three phases: the training phase, the testing phase, and the generalization phase. In the training phase, the participant was presented with a series of pairs of animals, the first always being a single adult animal and the second being either a baby animal or a group of animals. In the [a] condition, all of the obstruent final stems were affixed with the plural [-al] ending to create the symmetrical aCa frame for learning the spirantization rule (e.g. dolat ~ dolaθal). Similarly, in [i] condition training, all of the obstruent final stems were affixed with the diminutive suffix [-in] to establish the iCi frame (e.g. [jolid] ~ [joliðin]). The complementary suffix in each condition was introduced in the training phase via the filler words, which all contained an invariant stem final sonorant (e.g. monil ~ monilal/monilin). To preserve poverty of the stimulus, no obstruents appeared in asymmetric intervocalic frames during the training phase. The experimenter said the word, and the participant repeated the word aloud. If the participant produced the word incorrectly, the experimenter repeated the correct form of the word until the participant repeated it correctly.

After completing the training procedure for the entire training list (4 [t] final stems, 4 [d] final stems, 7 sonorant final stems), the participant was tested on the items they just learned. Going through the same list, the experiment said the word for the single animal, and the participant was instructed to produce the word for the group of animals or the baby animal. If they made a mistake or could not remember the correct word, the experimenter said the correct word and asked the participant to repeat that singular-affixed pair of words. This process continued, looping through the same list of words until the participant completed the entire list with three or fewer errors (≥ 80% accuracy).

After passing the first list, the participant was tested using the same procedure as above on a new list of words of the same form as those in the first training list (i.e. all of the items still preserved the poverty of stimulus condition.) This second list was included for two reasons. First, it eliminated potential false positive "learners" who may have been able to pass the first list due to brute force memorization. Second, it provided the participants a practice opportunity to apply the rule they were learning to novel items that conformed to the structure of the words they had been trained on before encountering non-conforming items in the generalization phase. Once again, the participant was looped through this list until they completed the entire list with three or fewer errors. Participants who could not pass through both training lists within 30 minutes were excused without continuing to the generalization phase.

After passing both training lists, the participant entered the generalization phase. In the generalization phase, the participant encountered a novel set of animals. Each animal now appeared as a triplet, beginning with the singular form followed by the two affixed forms. The order of the affixed forms was balanced throughout the list. The generalization phase consisted of the same 24x3 items regardless of condition (8 stems which conformed to the pattern learned in training, 8 stems containing the untrained vowel followed by a final, and 8 filler items.) The order of items as well as their associated animal was randomized for each participant. The generalization phase was the first time the participant encountered both non-conforming stem types and asymmetric intervocalic frames (i.e. aCi and iCa). The participant was instructed that they would go through this list only one time and that the experimenter would not give them feedback on their responses. As in the testing phase, the experimenter produced the word for the singular form appearing on the left half of the computer screen, the participant repeated the singular form, and then the participant produced the word for the picture that appeared on the right half of the screen. The singular form was provided by the experimenter before the participant produced each target item. A sample dialogue for one generalization phase triplet is shown below:

> *(Picture of a single adult crocodile on the left half of the screen, right half of screen blank)*
> Experimenter: [noron]
> Participant: [noron]

(Picture of a baby crocodile appears on the right half of the screen)
Participant: [noronin]
(Picture on the right half of the screen changes from a baby crocodile to a group of crocodiles)
Experimenter: [noron]
Participant: [noron]
Participant: [noronal]

6.3 Participants

The participants were UCLA undergraduates who received course credit for their participation. 48 people participated, and the responses of 39 participants were analyzed (n=20 for training on [i], n=19 for training on [a]). Nine total participants were excluded for not learning the pattern in the time allotted (7), not being a native English speaker (1), or experimenter error (1).

6.4 Data processing

The participants' responses were recorded and processed by either the author or an undergraduate research assistant who did not administer the task. For each obstruent-final stem in the generalization phase, the response was recorded as either a stop or fricative based on the researcher's perception. Any token in which the participant did not produce an obstruent in the position of interest was excluded (=57 tokens), resulting in 1152 tokens from 39 participants. Since the predictions of the Effort hypothesis and the Continuity hypothesis are most directly addressed using symmetric intervocalic frames, these were the tokens (582) upon which the initial analysis was performed. Given that the symmetric frame data do not conform to the predictions of either of these hypothesis, a subsequent analysis was then performed upon the full data set, using all 1152 tokens from both symmetric and asymmetric intervocalic frames.

7. Results and discussion of the symmetric frames

The mean spirantization rates with standard error for the symmetric frames in each condition are shown in Figure 1, and numeric values are given in Table 3 below:

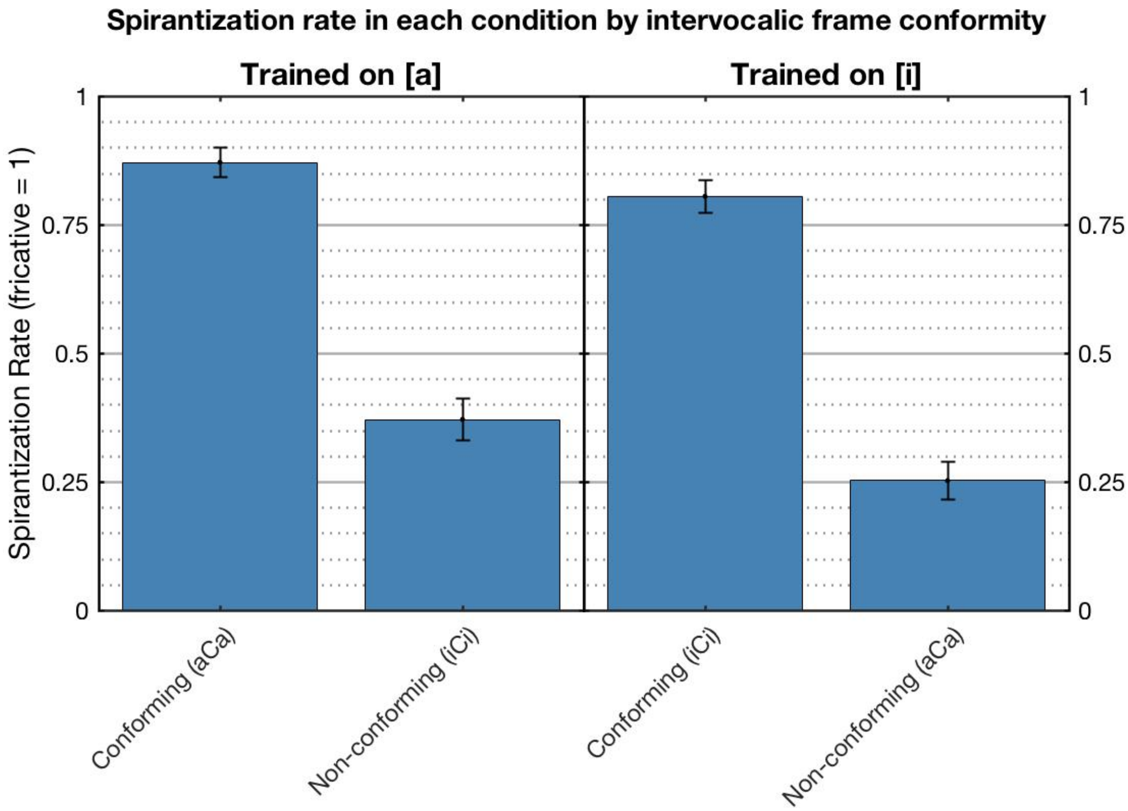

Figure 1; Mean spirantization rates (1=fricative) with standard error by condition and intervocalic frame conformity

	Conforming frame	Non-conforming frame
Trained on [a]	0.871 ± 0.028 (aCa)	0.371 ± 0.041 (iCi)
Trained on [i]	0.805 ± 0.031 (iCi)	0.252 ± 0.036 (aCa)

Table 3: Mean spirantization rates (1=fricative) with standard error by condition and intervocalic frame height

In both conditions, participants performed quite well applying the spirantization rule they were trained on to conforming items (i.e. aCa type words when trained on [a], iCi type words when trained on [i]). The high spirantization rate on conforming items in both conditions confirms that participants did in fact learn the pattern they were trained on.

In both conditions, participants were not as willing to spirantize words whose stem vowel differed in height from the training condition. This reluctance to generalize to novel intervocalic frame types goes against the predictions of the Continuity hypothesis.

The results of symmetric frames also do not conform to the predictions of the Effort hypothesis. Under the Effort hypothesis, the spirantization rates for both intervocalic frames in the Trained on [i] condition (right set of bars in Figure 1) are predicted to be the same. On the contrary, participants trained on [i] seem to be generalizing the pattern *even less* than those trained on [a].

Given that the results of the symmetric frames go against the predictions of both the Continuity hypothesis and the Effort hypothesis, we must search for an alternative hypothesis to account for the data. Additionally, since the Effort hypothesis is no longer a viable option, there is no justifiable reason to limit the scope of the data analysis to the symmetric frames (=582 tokens). As such, we now turn to the task of analyzing and modeling the full data set (=1152 tokens).

8. Results of the full data set (both symmetric and asymmetric intervocalic frames)

The numerical means with standard error for the full data set are shown in Table 4, and the results are plotted in Figure 2 below.

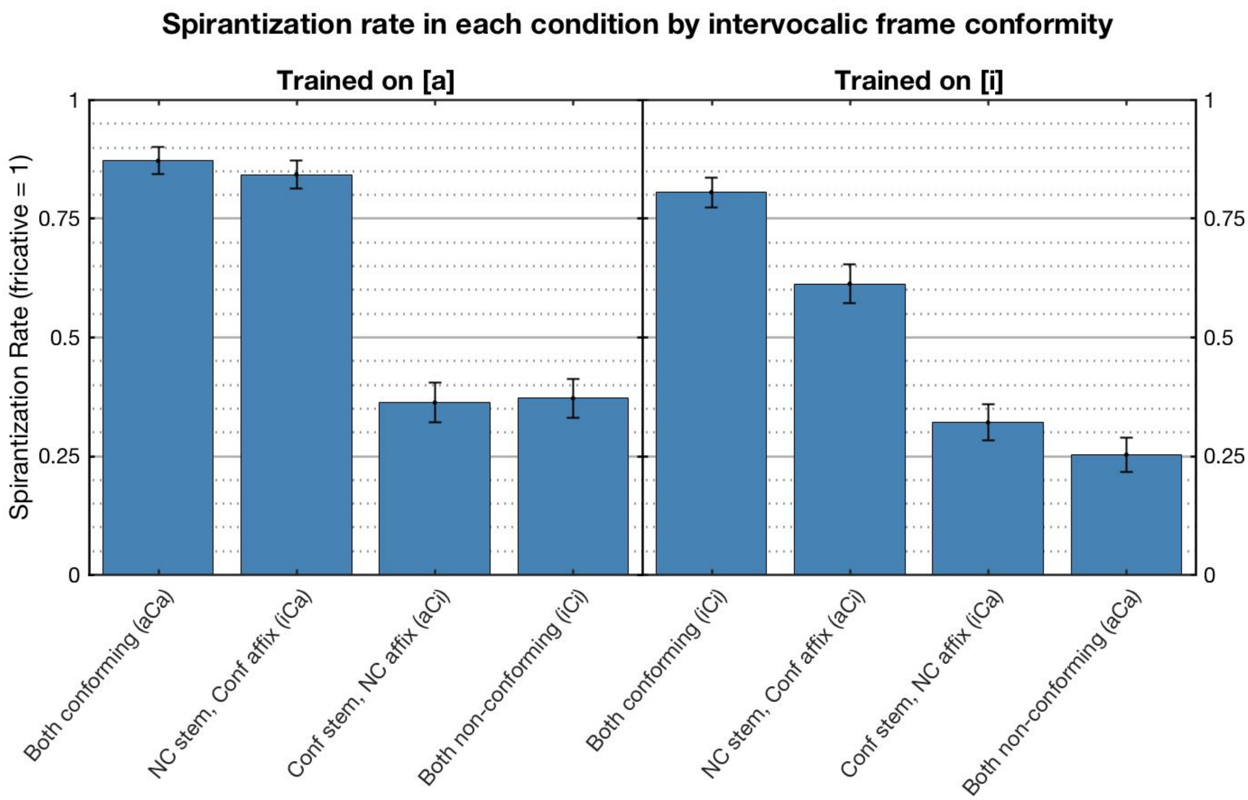

Figure 2: Spirantization rates(1=fricative) per condition by intervocalic frame conformity, with standard error, for all intervocalic frames

Trained on [a]	Conforming affix V	Non-conforming affix V
Conforming stem V	0.871± 0.028 (aCa)	0.362 ± 0.042 (aCi)
Non-conforming stem V	0.842 ± 0.030 (iCa)	0.371 ± 0.041 (iCi)
Trained on [i]		
Conforming stem V	0.805 ± 0.031 (iCi)	0.320 ± 0.039 (iCa)
Non-conforming stem V	0.612 ± 0.040 (aCi)	0.252 ± 0.036 (aCa)

Table 4: Mean spirantization rates (1=fricative) with standard error for all intervocalic frames

The data were analyzed with generalized linear mixed regression models using the glmer() function of the lme4 package (Bates et al. 2015) in R (R Core Team 2017). The dependent variable was continuancy (fricative = 1, stop = 0). The fixed effects were training condition (either [a] or [i]), conformity of the stem vowel, conformity of the affix vowel, and voicing. A fixed effect of experimenter (the author versus a research assistant naïve to the purpose of the experiment) was included in an earlier version of the model and was not significant (p=0.83). The model also included a random intercept of subject. A random intercept of item was also included in an earlier model but had a variance near zero and was therefore removed. A number of random slopes were tried, but none of them allowed the model to converge.

The fixed effects were encoded into the model using the deviation coding method (UCLA Statistics Consulting Center webpage, accessed June 2018). In deviation coding, the contrast coefficient for each level of a factor is the mean of that level minus the grand mean (the mean of the means of the dependent variable of each level of the categorical variable).

All four of the main effects were found to be significant. When the stem vowel conformed to the training condition, the probability of spirantization significantly increased (p=.001), with a contrast coefficient of 0.25. In the bar plot, this corresponds to the slight overall increase in height of bar one compared to bar two and bar three compared to bar four. The conformity of the affix vowel contributed a large effect in the direction of increased spirantization (contrast coefficient of 1.33) and was highly significant (p<<.0001). This main effect corresponds to the first two bars being much higher than the last two bars in each condition. The main effect of training vowel (condition) was also significant (p=.039), with the trained on [i] condition corresponding to a decrease in spirantization rate (contrast coefficient of -0.43). This means that overall, participants who were trained on [i] were less likely to spirantize regardless of intervocalic frame type than those who were trained on [a]. Finally, there was a highly significant (p<<.0001) main effect of voicing, with voiced obstruents being less likely to undergo spirantization than their voiceless counterparts (contrast coefficient of -0.48.) I am not sure what might be causing participants to spirantize less for voiced obstruents, which goes against typological and physiological evidence that voiced obstruents are more likely to spirantize than voiceless ones (Lavoie 2001). Since voicing is tangential to the main themes discussed in this paper, however, I will not attempt to speculate about it here.

There were also two significant fixed effect interactions. The interaction between stem vowel conformity and training vowel condition was found to be significant (p=.036), with conforming stems in the [i] condition being slightly more likely to spirantize (contrast coefficient of 0.166). This corresponds to the first and third bars being slightly higher in the [i] condition than expected when the relevant main effects are also taken into account. Finally, the interaction between the conformity of the suffix vowel and the training vowel condition was found to be approaching significance (p=.053), with the conforming suffix vowel [i] in the [i] condition being less likely to spirantize (contrast coefficient of -0.163). This corresponds to bars one and two in the trained on [i] condition being lower than expected. These two interactions conspire to explain the lowering of the second bar relative to the first bar in the [i] condition compared to the second bar in the [a] condition. The interaction between the conformity of the two vowels and the three way interaction of the conformity of the two vowels and training condition were both not significant.

A post-hoc series of Type II Wald chi-squared tests using the Anova() function in R showed that all of the values reported as significant were contributing meaningfully to the model (all p values < .05). A summary of the model is shown in Table 5 below.

Predictor	**Estimate**	**Standard Error**	***z*-value**	***p*-value**
Intercept	0.467	0.210	2.231	0.026
Stem V conformity –conf	0.252	0.079	3.189	0.001
Affix V conformity - conf	1.33	0.09	15.07	$< 2 \times 10^{-16}$
Training condition – [i]	-0.432	0.209	-2.065	0.039
Voicing - voiced	-0.484	0.079	-6.123	9.21×10^{-10}
Interaction – conf. stem V and conf. affix V	0.121	0.079	1.532	0.125
Interaction – conf. stem V and trained on [i]	0.166	0.079	2.095	0.036
Interaction – conf. affix V and trained on [i]	-0.163	0.084	-1.938	0.053
Interaction – conf. stem V, conf. affix V, train [i]	0.053	0.079	0.670	0.503

Table 5: Summary of the linear mixed effects model

Overall, these results suggest that participants trained in the [i] condition are more reluctant to spirantize overall, and they are less likely to generalize the pattern to the frames with low vowels than those trained on [a] to generalize to high vowel frames. These results are unexpected under both the Effort hypothesis and the Continuity hypothesis, but the evidence does align with the predictions of the Saliency hypothesis.

9. Discussion

The Saliency hypothesis claims that speakers are sensitive to the perceptual salience of potential sound changes. The more salient the change, the more resistant the speaker will be to incorporating that change into her grammar. This sensitivity to salience can be modeled in the grammar using *MAP constraints (Zuraw 2007, 2013). *MAP constraints are the grammatical manifestation of the speaker's knowledge of the P-Map (Steriade 2001b).

The working assumption here is that the shift from /ata/ to [aθa] is more perceptually salient than the corresponding change from /iti/ to [iθi]. This claim would be substantiated with a phonetic perceptibility experiment producing confusion matrix data (Miller & Nicely 1955) for the coronal obstruents of English in intervocalic position surrounded by both high and low vowels (i.e. iCi and aCa). To my knowledge, this type of experiment has not yet been performed, and I leave it as a direction for future research.

If the above assumption holds, the results of the spirantization learnability experiment are consistent with the Saliency hypothesis. The more salient sound change is learned by participants in the [a] training condition. Using a classic OT analysis, in the [a] condition speakers learn that the language contains a high-ranking markedness constraint against intervocalic stops (*VtV) that necessitates a relatively salient sound change (in this case, /ata/ → [aθa]). This markedness constraint must outrank the corresponding faithfulness constraint, *MAP(ata → aθa). Moreover, because (by hypothesis) [ata] → [aθa] is a more salient sound change than [iti] → [iθi], *MAP(iti → iθi) is by default ranked below *MAP(ata → aθa). Crucially, therefore, with no evidence to the contrary, the learner assumes that the markedness constraint *VtV outranks *both* faithfulness constraints since it outranks the one which has a higher default rank. As such, the learner should generalize the spirantization rule learned in the context of low vowels to the unfamiliar context between high vowels. Note that under a classic OT analysis, this should lead to categorical application of the rule in both environments rather than the gradient generalization the results suggest. The observed gradient application is easily handled, however, in a Harmonic Grammar framework such as MaxEnt (Smolensky 1986, Goldwater & Johnson 2003). Rather than assuming that the faithfulness constraints having a default ranking, we can incorporate them into the grammar with a default relative weighting. This method will be explored further in the next section.

In the [i] training condition, the same default ranking (or weighting) of *MAP(ata → aθa) >> *MAP(iti → iθi) is present in the grammar. In this condition, however, the learner only has evidence that the markedness constraint *VtV must dominate the lower ranking faithfulness constraint, *MAP(iti → iθi). In the absence of evidence about the ranking of *VtV relative to the *MAP(ata → aθa), the learner has no justification to promote *VtV above the higher ranked faithfulness constraint. Therefore, this hypothesis predicts lower overall rates of generalization of the spirantization rule to the low vowel frame when the learner is trained on [i]. This pattern of asymmetric generalization corresponds precisely with the results of the experiment detailed above.

Given this theoretical schema, we can seek to model the results of the experiment implementing the constraints specified by the Saliency hypothesis. It is to this endeavor that we now turn.

10. A MaxEnt model of learning using constraint weight biases

The Maximum Entropy (Smolensky 1986, Goldwater & Johnson 2003) variety of Harmonic Grammar provides a viable framework to model the artificial grammar learning results primarily for two reasons. First, MaxEnt has a mechanism for implementing constraint weight biases such as those posited by the Saliency hypothesis. Secondly, MaxEnt is designed to model probabilistic outputs such as the responses of the participants in the experiment.

MaxEnt operates by assigning weights to the constraints of the grammar that maximize the log probability of the data set. On its own, however, this procedure does not account for a constraint's "default weight," which may correspond to a UG bias or to the nature of the constraint definition. For example, since *MAP constraints are the formal implementation of P-map relations in the grammar, the default weights for these constraints should be derivable in some way from the perceptual distance of the correspondence relations they are mapping. Wilson (2006) makes an attempt at formalizing perceptual distance into constraint weight biases using the Generalized Context Model of classification (Nosofsky 1986). This model employs a combination of phonetic features (both discrete and continuous) and confusion matrix data to produce a quantitative version of the P-map. These constraint weight biases can then be incorporated into the MaxEnt evaluation procedure using a Gaussian prior penalty term (Chen & Rosenfeld 1999).

The full equation for MaxEnt objective function including the penalty term is as follows:

$$\sum_{i=1}^{m} \log(P(y_i|x_i)) - \sum_{j=1}^{n} \frac{(w_j - \mu_j)^2}{2\sigma_j^2}$$

The first term of this equation is the sum of the log probability of every input-output pair in the data set. The second term is the Gaussian prior penalty term. This term is composed of the square of the difference of the constraint weight (w_j) minus that constraint's expected value (μ_j), divided by twice the square of that constraint's propensity to deviate from its expected weight (σ_j). Since it appears in the denominator of the penalty term, a larger σ_j corresponds to a decreased penalty for deviating from the constraint's expected weight. If the constraint has no default weight bias, $\mu_j = 0$. By adjusting a constraint's μ, therefore, we can formally incorporate default constraint weight biases into the grammar.

We can now investigate how well the theoretical predictions of the Saliency hypothesis conform to the artificial grammar learning results using MaxEnt. First, we assign μ and σ values for each of the constraints in the grammar. Next, the MaxEnt objective function determines the weights of the grammar using the same training data shown to participants as the training input. Using these weights, MaxEnt determines the probability distribution of the responses to novel items.

Following the work of White (2017), I adjusted the μ values for each *MAP constraint and used the same value for σ for all of the constraints ($\sigma = 2$). In the absence of the most directly applicable confusion matrix data (comparing [ata] → [aθa] with [iti] → [iθi]), I based the constraint weight biases on the consonant confusion matrix data of Wang & Bilger (1973). The relevant data are summarized in Table 6 below:

Vt heard as Vθ	62	tV heard as θV	30
Total Vt tokens	1721	Total tV tokens	1713
% Vt heard as Vθ	3.6%	% tV heard as θV	1.8%
%Vt/tV heard as Vθ/θV (μ_0)	**2.6%**		

Table 6: t/θ confusion matrix data from Wang & Bilger (1973)

The following method of assigning constraint weight biases is slightly ad hoc, but it is intended to make reasonable assumptions and is somewhat rooted in perceptual data as opposed to pulling random numbers out of a hat in order to suit the analysis. First, I set the composite [t]→[θ] confusability percentage as μ_0 (= 2.6) based on Wang & Bilger's data. This was the μ assigned to the generic *MAP(t→θ), which was active in the analysis to prevent stem final stops from surfacing as fricatives ([tanat]/*[tanaθ]). Next, since the intervocalic frames have acoustic cues available from both the preceding and following vowel, I set the bias for the [iCi] frame as $2\mu_0$ (=5.2). Finally, since the Saliency hypothesis is working under the assumption that [aCa] is a more perceptually salient frame than [iCi], I incorporated a 10% bump and set the bias for the [aCa] frame to be $2.2\mu_0$ (= 5.72). The constraint weight bias for the markedness constraint *VtV was set to μ = 0.

With a set of constraint weight biases in place, the MaxEnt model is ready to learn the relevant training data. Crucially, the constraint weight biases must be the same for both the training on [a] and training on [i] MaxEnt models because the perceptibility of the respective sounds is the same regardless of the training condition. The analysis was done in Microsoft Excel using the Solver add-on. The model was trained on 48 observations of the conforming stem/affixed form (e.g. [tanat], [tanaθal]) to reflect the fact that a typical participant went through the first list about three times and the second list twice. The weights found by the MaxEnt model for both conditions are shown in Table 7 below:

	Trained on [a]	Trained on [i]
*VtV (μ = 0)	4.38	4.15
*MAP(ata → aθa) (μ = 5.72)	1.34	5.72
*MAP(iti → iθi) (μ = 5.2)	5.20	1.05
*MAP(t → θ) (μ = 2.6)	4.13	4.13

Table 7 Constraint weights for a biased MaxEnt grammar trained on either the [a] or [i] condition

Note that as expected, the non-conforming faithfulness constraint in each condition was left at its default weight since the model had no reason to shift it and endure the corresponding penalty. Given these weights, let us now investigate how the models' predictions compare to participants' responses on the novel items. A bar plot comparing the models' predictions to observed responses for each of the intervocalic frames in each condition is shown in Figure 3 below. For the asymmetric frames, there is no directly applicable *MAP constraint, and so faithfulness violations must be assigned through some combination of the two available *MAP constraints. Since the following vowels seems to be the primary driving force in determining spirantization application rates, I assigned a violation of 0.67 to the *MAP constraint associated with the following vowel, and a violation of 0.33 to the *MAP constraint of the preceding vowel for a total *MAP violation of 1. For example, [tiniθal] would incur a 0.67 violation to *MAP(ata → aθa) and a 0.33 violation to *MAP(iti → iθi).

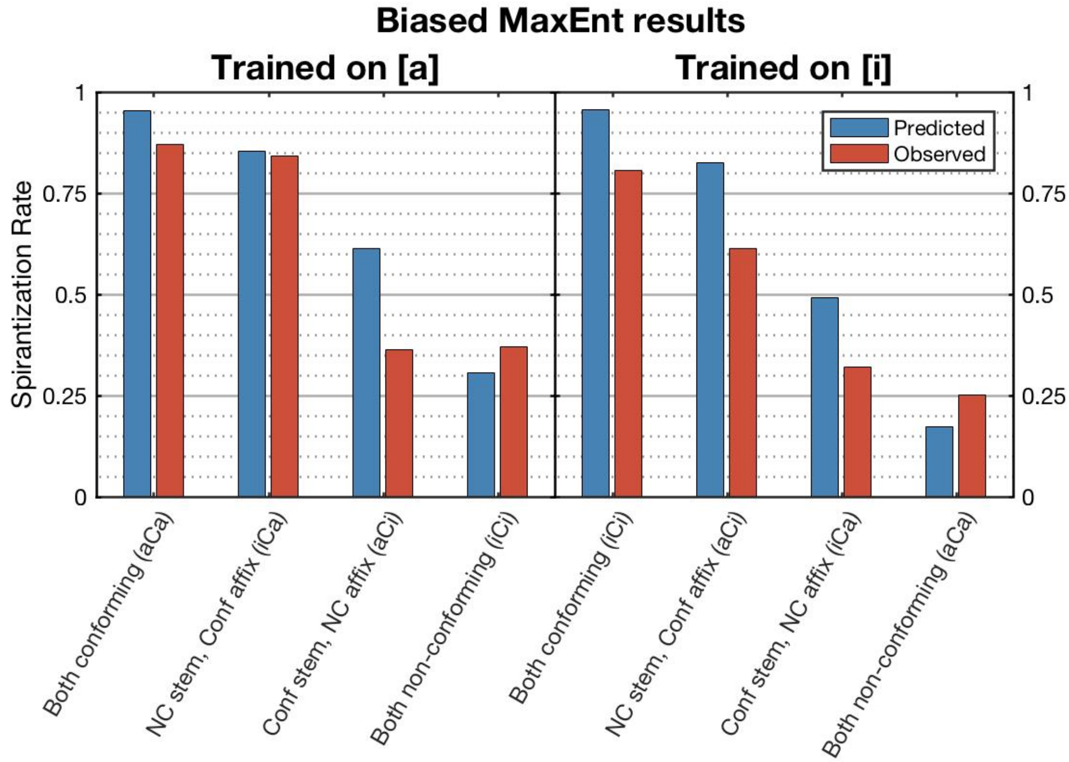

Figure 3: Spirantization rates predicted by a biased MaxEnt grammar (left, blue) trained on each condition compared to the spirantization rates observed in the experiment (right, red)

As the plots demonstrate, the Saliency biased MaxEnt models are a very close fit to the observed data. The overall predicted rate of spirantization is higher for the high vowel frame when trained on the low vowel (set of bars fourth from the left, =0.306) than for the low vowel frame when trained on the high vowel (rightmost set of bars, =0.172). Although the models underpredict spirantization in both of these cases, the underprediction is consistent across both conditions (~0.7 - 0.8). In fact, the MaxEnt models predict slightly more extreme spirantization values for all four of the symmetric frame/training vowel combinations. This could be a smoothing effect of the artificial grammar learning experiment or perhaps noise from the experimental data. The predictions corresponding to the asymmetric frames are not quite as close a match in three of the four cases, but I attribute this to the patchwork implementation of constraint violations in the absence of a directly applicable *MAP constraint.

To reinforce the goodness of the MaxEnt model fit using saliency-motivated constraint weight biases, we can compare the biased model to an unbiased MaxEnt model (all $\mu = 0$, and $\sigma = 10,000$ to avoid overfitting) trained on the same data. The predictions of the unbiased model for novel items can be seen in the bar plots in Figure 4 below. In both conditions, the model sets the weights of the faithfulness constraints to zero and the markedness constraint receives a very high weight. This causes the model to vastly overpredict the spirantization rate for all four frame/training vowel combinations. Clearly, the saliency-biased MaxEnt model is accounting for the data much more convincingly than the unbiased MaxEnt model.

Figure 4: Spirantization rates predicted by an unbiased MaxEnt grammar (left, blue) trained on each condition compared to the spirantization rates observed in the experiment (right, red)

11. Conclusions

The goal of this paper was to compare two different theoretical frameworks that provide a synchronic explanation for intervocalic lenition, using empirical evidence. The two frameworks make differing predictions regarding the role of vowel height in lenition processes such as intervocalic spirantization. The Effort hypothesis (Kirchner 1998, 2004) predicts that lenition should happen more frequently between low vowels, whereas the Continuity hypothesis (Kingston 2008, Katz 2016) predicts that vowel height should not be a factor in determining the rate of lenition.

To test these predictions, an artificial grammar learning experiment (Wilson 2006) was conducted. There were two conditions, one in which participants were trained on the rule /ata/ → [aθa] and the other in which participants were trained on /iti/ → [iθi]. When participants encountered words that did not conform to the pattern they had learned, they generalized the spirantization rule at different rates depending on their training condition. Those who were trained on [a] were more willing to generalize to spirantizing in the high intervocalic frame than those trained on [i] to the low intervocalic frame. These results seem to be best explained by a third hypothesis, perceptual saliency (Steriade 2001a, 2001b; Fleischhacker 2005). Participants who learn the sound change in a more salient position are more willing to generalize to a less salient position than vice versa. The Saliency hypothesis was then modeled using constraint weight biases in MaxEnt and was shown to fit well to the results of the spirantization experiment. The claims made by the Saliency hypothesis depend on the relative perceptibility of spirantization between low vowels versus spirantization between high vowel. Based on the results of this experiment, the Saliency hypothesis predicts that a perceptual confusability study comparing [ata] → [aθa] with [iti] → [iθi] will find greater confusability in the high vowel frame. This phonetic perception study is left for future research.

It is interesting to note that despite the design of the experiment as a production-based task, the results were best explained using perceptual similarity. Though the experiment was designed to explore lenition hypotheses, the Saliency hypothesis is agnostic as to the driving force behind lenition. As such, several questions regarding the status of lenition processes in synchronic phonology remain. The results do indicate, however, that neither the Effort hypothesis nor the Continuity hypothesis is a satisfactory mechanism to explain the synchronic lenition of the experiment presented here.

Acknowledgements

This project would not have succeeded without the help and guidance of many people. In particular, I would like to thank my MA committee chair Bruce Hayes and my committee members Kie Zuraw and Pat Keating for their encouragement and many helpful suggestions, as well as Jesse Harris, Eleanor Glewwe, and the UCLA Statistics consultants for their assistance in interpreting my models. I would also like to thank my undergraduate research assistants Lakenya Riley, Arjun Gananathan, and Derek Zhang as well as the UCLA Psychology Subject Pool participants. For their helpful feedback and innovative suggestions, I thank the UCLA Phonology Seminar as well as my fellow graduate students, particularly Travis Major, Canaan Breiss, Z.L. Zhou, Connor Mayer, Jesse Zymet, and Jeremy Steffman. Thank you!

Appendix A: Full list of experiment stimuli

Train [a] list 1

picture	word
alpaca	monil
alpaca baby	monilin
alpaca herd	monilal
bat	torat
bat herd	toraθal
beaver	jarol
beaver baby	jarolin
beaver herd	jarolal
horse	tanad
horse heard	tanaðal
monkey	larat
monkey herd	laraθal
flamingo	jorim
flamingo baby	jorimin
elephant	ralad
elephant herd	ralaðal
buffalo	dilam
buffalo baby	dilamin
cow	dolat
cow herd	dolaθal
owl	ronam
owl herd	ronamal
kangaroo	jinad
kangaroo herd	jinaðal
cheetah	nilon
cheetah baby	nilonin
loon	dorad
loon herd	doraðal
rabbit	tinat
rabbit herd	tinaθal
eagle	tiron
eagle herd	tironal

Train [a] list 2

picture	word
camel	milom
camel herd	milomal
camel baby	milomin
pig	ramad
pig herd	ramaðal
moose	jomat
moose herd	jomaθal
dog	lolin
dog baby	lolinin
crab	tilat
crab herd	tilaθal
chicken	tolir
chicken baby	tolirin
bison	jolad
bison herd	jolaðal
cat	nomat
cat herd	nomaθal
antelope	molor
antelope baby	molorin
bobcat	lirad
bobcat herd	liraðal
grizzly	tonil
grizzly baby	tonilin
narwhal	tarat
narwhal herd	taraθal
peacock	momad
peacock herd	momaðal
rhino	limor
rhino baby	limorin
yak	tarad
yak herd	taraðal
vulture	nanat
vulture herd	nanaθal
jellyfish	torin
jellyfish herd	torinal

Train [i] list 1

picture	word
alpaca	monar
alpaca baby	monarin
alpaca herd	monaral
bat	torit
bat baby	toriθin
beaver	jarol
beaver baby	jarolin
beaver herd	jarolal
horse	tanid
horse baby	taniðin
monkey	rolit
monkey baby	roliθin
flamingo	riram
flamingo herd	riramal
elephant	nirid
elephant baby	niriðin
buffalo	dilam
buffalo baby	dilamal
cow	domit
cow herd	domiθin
owl	ronam
owl baby	ronamin
kangaroo	jonid
kangaroo baby	joniðin
cheetah	nilon
cheetah herd	nilonal
loon	darid
loon baby	dariðin
rabbit	dinit
rabbit baby	diniθin
eagle	tiron
eagle baby	tironin

Train [i] list 2

picture	word
camel	milam
camel herd	milamal
camel baby	milamin
pig	ramid
pig baby	ramiðin
moose	jomit
moose baby	jomiθin
dog	lolin
dog herd	lolinal
crab	talit
crab baby	taliθin
chicken	tolir
chicken herd	toliral
bison	jolid
bison baby	joliðin
cat	nimit
cat baby	nimiθin
antelope	molor
antelope herd	moloral
bobcat	larid
bobcat baby	lariðin
grizzly	tonan
grizzly herd	tonanal
narwhal	tarit
narwhal baby	tariθin
peacock	momid
peacock baby	momiðin
rhino	lamor
rhino herd	lamoral
yak	tanid
yak baby	taniðin
vulture	nanit
vulture baby	naniθin
jellyfish	rolam
jellyfish baby	rolamin

Generalization list (randomized for each subject)
noron
torid
tolat
rolid
jonit
nojad
lalad
nalan
lanit
momor
nilid
naril
ranol
rilat
lirad
dolit
lalon
lamat

Appendix B: Iterations of Experimental Methodologies

From conception to the final version, the experimental task design went through several iterations. In any artificial grammar learning paradigm, the experimenter must balance the difficulty of the task with the risk of either no learning or explicit learning of the alternation, as visualized below:

Less difficult ⟷ More difficult

Risk: Explicit Learning | Risk: No learning

I began the design of my task to the extreme right of the spectrum and found that participants were not learning the alternation in any meaningful way. I will now detail some of the many alterations I made to the experimental paradigm so that others who wish to conduct a similar experiment might not fall victim to the same pitfalls.

Version 1: A computerized task performed remotely over the internet using the Appsobabble platform (developed by Henry Tehrani). Participants were instructed to wear headphones. There were two training/testing blocks (train, test, train, test) followed by the generalization phase. In the training block, participants saw an animal picture, heard a pre-recorded stimulus, and were instructed to repeat the word they heard aloud. The training block was randomized such that the participant likely never saw a full morphological paradigm together (e.g. baby panda followed by herd of horses followed by single flamingo). In the testing phase, the participant heard a sequence of two words which differed only in the continuancy value of the obstruent (e.g. tanat –

tanaθ) and were asked to choose via key press whether the first or second word they heard matched the language they had learned during the training phase. After doing both the training and testing block once, they repeated both again regardless of performance. The generalization phase was of the same form as the testing phase.

Version 2: Same as V1 but added a third training/testing block to give participants another chance to learn the rule before generalization.

Version 3: Reduced the total number of items from 24x3 to 20x3, thinking perhaps fewer items would be slightly less information for participants to process

Version 4: Brought participants into the lab rather than performing the task online. This was done to improve the acoustic environment and eliminate potential distractions. (It was a relatively long task for a pure online experiment, ~35-45 minutes)

Version 5: Reduced target items to exclude asymmetric VCV frames, so as to not unduly privilege the preceding vowel in exploring the effort hypothesis. This reduced the number of items in the training phase from 20x3 to 20x2.

Version 6: Removed total randomization. Participants now saw a randomized paradigm for each item (baby pig/single pig/herd of pigs, single horse/herd of horses/baby horse, etc.)

Version 7A: Eliminated the voicing contrast so participants only saw voiceless words.

Version 7B: Updated the instructions of V6 to encourage participants to pay more attention to the alternations: "You will see pictures of a single animal, a group of animals, and a young animal. Try to learn how the word for a group of a particular animal or a young animal relates to the word for that single animal."

Version 8: Updated V7B to allow participants to replay the stimulus by clicking on the picture again.

All of the above computerized iterations were piloted with at least four participants and none showed any meaningful learning of the alternation. After these eight versions, I switched to the in-person training model that was eventually used to run the full experiment. The only significant change between the first in-person pilot and the final version was the introduction of the intermediate conforming list. This second list was included for two reasons. First, it eliminated potential false positive "learners" who may have been able to pass the first list due to brute force memorization. Second, it provided the participants a practice opportunity to apply the rule they were learning to novel items that conformed to the structure of the words they had been trained on before encountering non-conforming items in the generalization phase.

There are a few reasons I believe the shift to the in-person paradigm was so much more successful than the computerized versions of the task. First, the in-person version eliminated the need for the participant to ever encounter wrong forms. During the testing phase of the computerized version, half of the forms that the participant hears are incorrect. Artificial grammar learning experiments hinge on participants mastering an alternation in a very short amount of time (30 minutes is nothing compared to the amount of time we spend learning a natural language), and so every minute they spend during the task is precious learning time. Even during the "testing" phase, there is likely still lots of learning happening and hearing so many incorrect forms is undoubtedly a hindrance in accomplishing successful learning. In the in-person version, however, the participant never hears an incorrect form unless she produces it herself.

Second, the in-person version allows for immediate feedback and correction. Some computerized platforms may also have this feature, but getting immediate feedback seems like an important stage in the learning process. The in-person version allows for the participant to continue to repeat the item until she produces it correctly.

Third, my impression is that participants are willing to work harder at the task when they are interacting with another human being rather than with a computer. The field of psychology probably has much more to say regarding the differences between person-to-person versus person-machine interactions, but I noticed that most participants really seemed to want to do well on the task even if they thought it was hard. After finishing the task, many more of those who completed the in-person version volunteered that they had enjoyed the task as compared to those who completed a computerized version.

Of course, the in-person version is not without its own flaws. Most notably, there is the risk that the experimenter will subconsciously influence the responses of the participant. There is also more risk of variability between trials, say if the experimenter makes a mistake in pronouncing one of the items. There is also the drawback that data analysis is not as automatable in a production task as compared to a forced choice task.

To mitigate the drawbacks as much as possible while maximizing learning potential, and by extension the artificial grammar learning paradigm itself, I propose the following methodology. Training should take place in person with a trained experimenter blind to the purpose of the task. After the participant has met the pre-determined threshold criterion for learning, the experimenter leaves the room and the participant performs a computerized generalization task. This should probably still be a recorded production task so as to still avoid the participant hearing wrong forms for the sake of testing. For example, in the case of the spirantization experiment discussed in this paper, this would have looked like the participant seeing a picture of a single animal, hearing the word for that animal, and then seeing the diminutive or plural version of that animal and being asked to produce the correct word. Data analysis for this task would have the same time requirements as the completely in-person version of the task.

References:

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Chen, S. F., & Rosenfeld, R. (1999). *A Gaussian prior for smoothing maximum entropy models* (No. CMU-CS-99-108). CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

Dixon, R. M. W. (1991). Mbabaram, 348-402. *The handbook of Australian languages*, *4*.

Doke, C. (1957) A Textbook of Southern Sotho Grammar, Longmans, Green and Co., London.

Fleischhacker, H. A. (2005). *Similarity in phonology: Evidence from reduplication and loan adaptation* (Doctoral dissertation, University of California, Los Angeles).

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The journal of the Acoustical Society of America*, *101*(6), 3728-3740.

Goldwater, S., & Johnson, M. (2003, April). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory* (Vol. 111120).

Grammont, M. (1933) Traité de Phonetique, Delgrave, Paris

Gurevich, N. (2005). Lenition and Contrast. New York: Routledge.

Introduction to SAS. UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/ (accessed June 16, 2018).

Katz, J. (2016). Lenition, perception and neutralisation. *Phonology*, *33*(01), 43-85.

Keating, P. A., Lindblom, B., Lubker, J., & Kreiman, J. (1994). Variability in jaw height for segments in English and Swedish VCVs. *Journal of Phonetics*, *22*(4), 407-422.

Keating, P., Cho, T., Fougeron, C., & Hsu, C. S. (2004). Domain-initial articulatory strengthening in four languages. *Phonetic interpretation: Papers in laboratory phonology VI*, 143-161.

Keating, P. A. (2006). Phonetic encoding of prosodic structure. *Speech production: Models, phonetic processes, and techniques*, 167-186.

Kingston, J. (2008). Lenition. In *3rd Conference on Laboratory Approaches to Spanish Phonology* (pp. 1-31). Cascadilla Proceedings Project.

Kirchner, R. (1998). An effort-based account of consonant lenition. *Unpublished doctoral thesis). University of California at Los Angeles, Los Angeles, USA*.

Kirchner, R. (2004). Consonant Lenition. In Hayes, B., Kirchner, R., & Steriade, D. (Eds.). (2004). *Phonetically based phonology*. Cambridge University Press.

Lavoie, L. M. (2001). *Consonant strength: Phonological patterns and phonetic manifestations*. Routledge.

Leal, D. (1972) Chitwan Tharu Phonemic Summary, Kirtipur: Summer Institute of Linguistics and Institute of Nepal Studies, Tribhuvan University.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry*, *8*(2), 249-336.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338-352.

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

Prince, A., & Smolensky, P. (2008). *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory* (No. CU-CS-321-86). Colorado University at Boulder Dept of Computer Science.

Steriade, D. (2001a). Directional asymmetries in place assimilation: a perceptual account. *The role of speech perception in phonology*, 219-250.

Steriade, D. (2001b). The phonology of perceptibility effects: the P-map and its consequences for constraint organization. *Ms., UCLA*.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, *54*(5), 1248-1266.

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition*, *130*(1), 96-115.

White, J. (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language*, *93*(1), 1-36.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, *30*(5), 945-982.

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: corpus and survey evidence from Tagalog infixation. *Language*, *83*(2), 277-316.

Zuraw, K. (2013). *MAP constraints. Unpublished manuscript, UCLA.
www.linguistics.ucla.edu/people/zuraw/dnldpprs/star_map.pdf