Similarity and bias in phonological learning

Colin Wilson
UCLA

Rumelhart Prize Symposium
CogSci 2005, July 22

Overview

- What role does **phonetic substance** play in phonological cognition?
- Results from language game experiments on velar palatalization (k→t͡ʃ "ch", g→d͡ʒ "j") show **asymmetric generalization patterns**
- Modeling with GCM and Maxent supports the claim that substance functions as a **bias** (or **prior**) on phonological learning

The role of substance

Limits on phonological cognition

- Formal universals
  Architecture (e.g., rules vs constraints)
  Formal language theory (e.g., finite-state phon)

- **Substantive universals**
  Does the phonetic content of phonological symbols also impose limits on phon cog?

Yes	Chomsky & Halle 1968, Hayes et al. 2004
Ltd	Halle 2005: distinctive features only
No	Blevins 2004, Ohala 1995, Hale & Reiss 2000

Typological evidence for substance

- Certain phonological patterns are found in many, genetically-diverse languages while others are rare or unattested.

  Ex. Word-final devoicing (Catalan, German, Ingush, Turkish, Wolof, . . .) vs. word-final voicing (Lezgian)

- The widely-attested phonological patterns can often be understood in terms of ease of articulation and/or perceptual distinctiveness

  Ex. Reduction of perceptual cues for the t/d contrast word-finally facilitates neutralization (Steriade 1999)

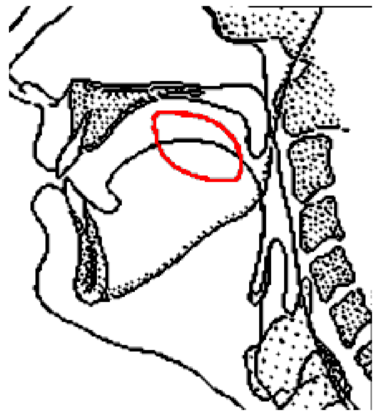
Case study: Velar palatalization

- Velar pal refers here to the change from a velar stop (k g) to a palatoalveolar affricate (t͡ʃ d͡ʒ) in certain vowel environments

- Components of the case study
  1. Phonetic properties of velar stops
  2. Typological generalization
  3. Steriade's law of similarity

1. Phonetic properties: Articulation

Velar stop + Vowel Coarticulation



k̟ / __i (most fronted)

k / __e

k̠ / __ɑ (least fronted)

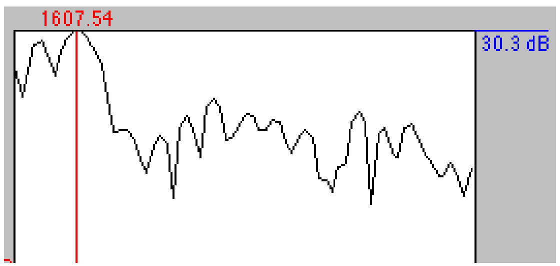
(pics from http://www.umanitoba.ca/faculties/arts/linguistics/russell)

References on velar fronting: Butcher & Tabain 2004, Keating & Lahiri 1993, Ladefoged 2001; see also Fletcher 1997, Fowler & Brancazio 2000 on V-to-V coarticulation across velars

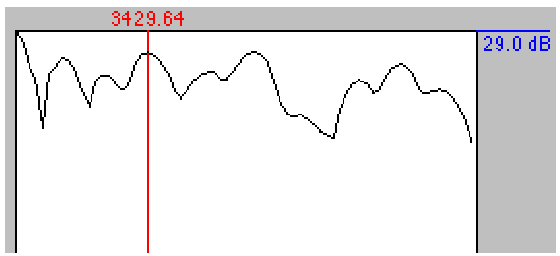
1. Phonetic properties: Acoustics

Articulatory fronting in the environment of front vowels gives rise to acoustic similarity of velar stops and palatoalveolar affricates

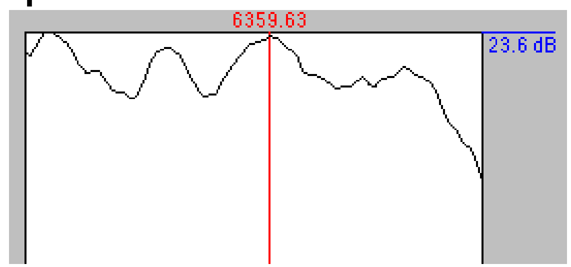
k̠ /__ ɑ (from ’kɑgə)



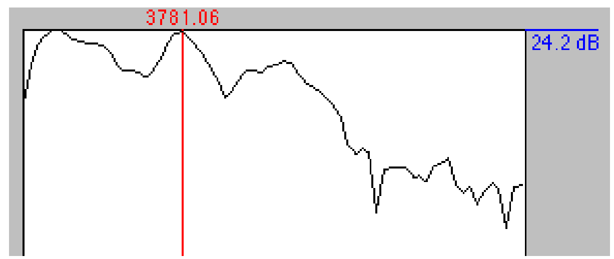
k / __ e (from ’kegə)



k̟ /__ i (from ’kigə)



Burst spectra for velar stops and the palatoalveolar affricate t͡ʃ (from ’t͡ʃigə)



Details of acoustic measurements

- Guion 1996, 1998 measured the peak frequency (kHz) in the average spectrum of the burst/aspiration (512 FFT points, +6dB preemphasis)

	back	front	high front
k / g	2.25	3.5	4.0
t͡ʃ / d͡ʒ	4.25	4.25	4.5

- These results (and measurements on the stimuli for the present experiments) show that **velar stops are more similar to palatoalveolar affricates before front vowels, especially high front vowels**

1. Phonetic properties: Perception

Acoustic similarity of velar stops and palato-alveolar affricates is correlated with rate of velar → pal errors in identification (Guion 1996, 1998)

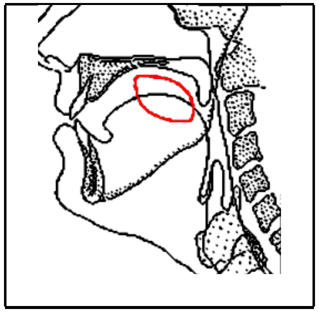
	Response							
Stimulus	[ki]	[tʃi]	[gi]	[dʒi]	[kɑ]	[tʃɑ]	[gɑ]	[dʒɑ]
[ki]	43	35	10	12				
[gi]	4	4	71	21				
[kɑ]					84	13	3	0
[gɑ]					4	0	87	9

2. Typological generalization

If velar palatalization ($k \rightarrow \widehat{t\text{ʃ}}$, $g \rightarrow \widehat{d\text{ʒ}}$) applies before a vowel V, then it also applies before all vowels that are more front than V

Ex. velar pal before e $\Rightarrow$ velar pal before i

Ex. velar pal before æ $\Rightarrow$ velar pal before e,i



Typological studies by Bhat 1978, Chen 1972, 1973, Guion 1996, 1998, Neeld 1973 document this generalization; see also Maddieson & Precoda 1992 on phonotactic reflexes

3. Steriade's law of similarity

- The cognitive system privileges alternations involving perceptually-similar sounds
  $\Delta_P(x, y) < \Delta_P(x, z)$
  $\Rightarrow \quad x \rightarrow y$ is preferred over $x \rightarrow z$

- The general law forms part of an explanation of the typological implication on velar pal

articulation	acoustics	perception	phonology
$V_1 > V_2$	$\widehat{t\int} \geq kV_1 > kV_2$	$\Delta_P(k,\widehat{t\int} \mid V_1)$	$k \rightarrow \widehat{t\int} \mid V_2$
(frontness)	(spectral peak)	$< \Delta_P(k,\widehat{t\int} \mid V_2)$	$\Rightarrow k \rightarrow \widehat{t\int} \mid V_1$

Evolutionary alternative

- An alternative explanation for the typological generalization would invoke (non-cognitive) diachronic change or "evolution"

- Greater acoustic/perceptual similarity of palatoalveolars and velars before more front vowels could make $k \rightarrow \widehat{t\int}$ __i more likely as a sound change ("error pattern") than $k \rightarrow \widehat{t\int}$ __e

See Blevins 2004, Blevins & Garrett 2004, Hale & Reiss 2000, Ohala 1992, 1995 on evolutionary phonology

Proposal

- Knowledge of substance functions as a bias (not absolute restriction) on phon grammars
- Weaker version of phonetically based phonology that avoids the empirical problems noted by Blevins and others
- Expect bias to be revealed most strongly when the input to the learner is impoverished

Language game experiments

Language games

- Language games are naturally-occurring phenomena that systematically alter the pronunciation of words (see Bagemihl 1995)
- Experiments reported here use artificial language games with impoverished input
- The measure of interest is how participants **generalize** from the input to a new vowel context ("poverty of the stimulus method")

Experiment 1: High vs Mid

- High exposure

kimə . . . t͡ʃimə $^{\times 8}$		kapə . . . kapə $^{\times 6}$
pilə . . . pilə	pebə . . . pebə	parə . . . parə

- Mid exposure

	kenə . . . t͡ʃenə $^{\times 8}$	kapə . . . kapə $^{\times 6}$
pilə . . . pilə	pebə . . . pebə	parə . . . parə

- Testing (both conditions; also included fillers)

kimə . . . ? $^{\times 8}$	kenə . . . ? $^{\times 8}$	kapə . . . ? $^{\times 6}$
kisə . . . ? $^{\times 8}$	kezə . . . ? $^{\times 8}$	kavə . . . ? $^{\times 6}$

Qualitative predictions

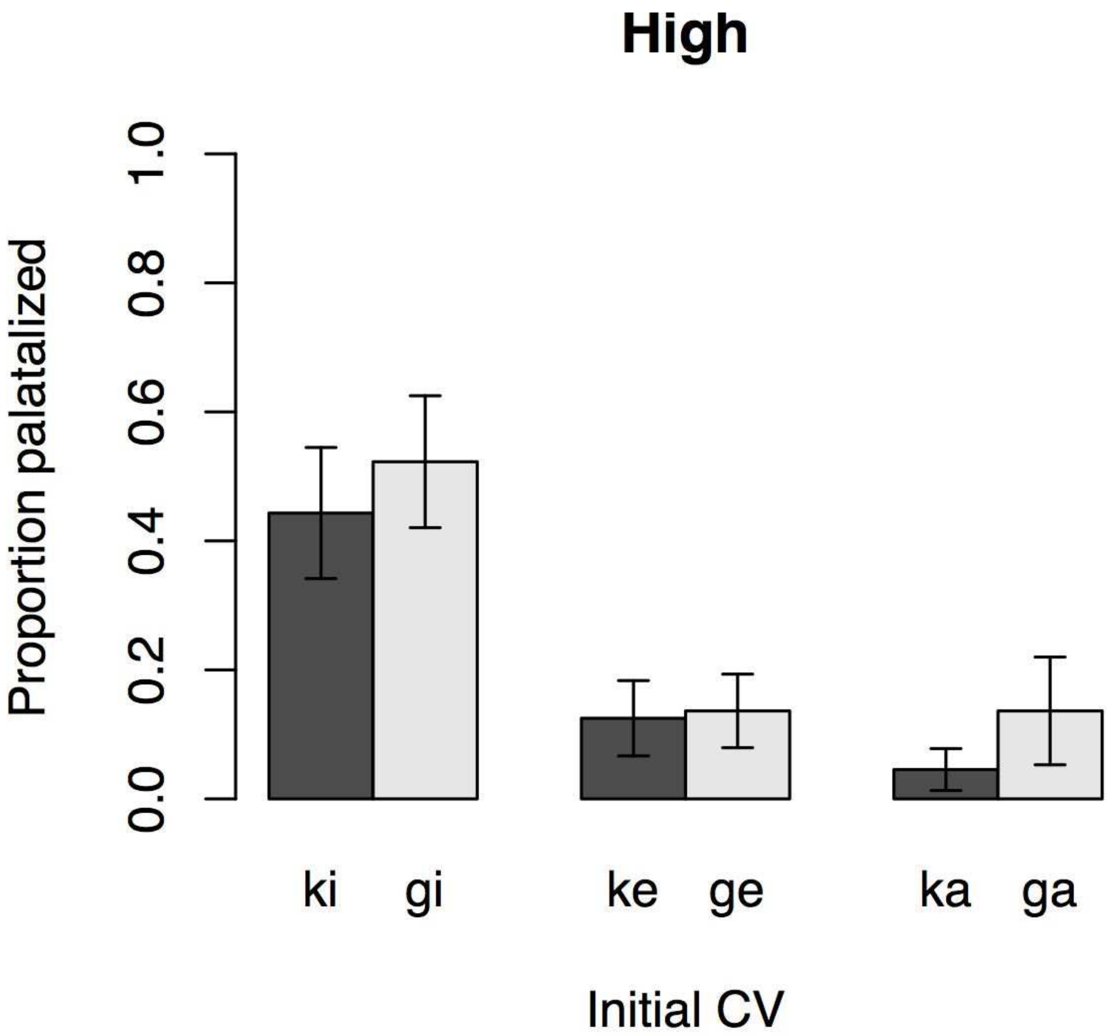
- If Steriade's law forms part of a bias (or prior) on phonological learning, then:

+ participants in the Mid group (k $\longrightarrow$ $\widehat{t\int}$ / __ e) should generalize to the novel context (__i)

– but do not expect the same degree of generalization in the High group (k $\longrightarrow$ $\widehat{t\int}$ / __i)

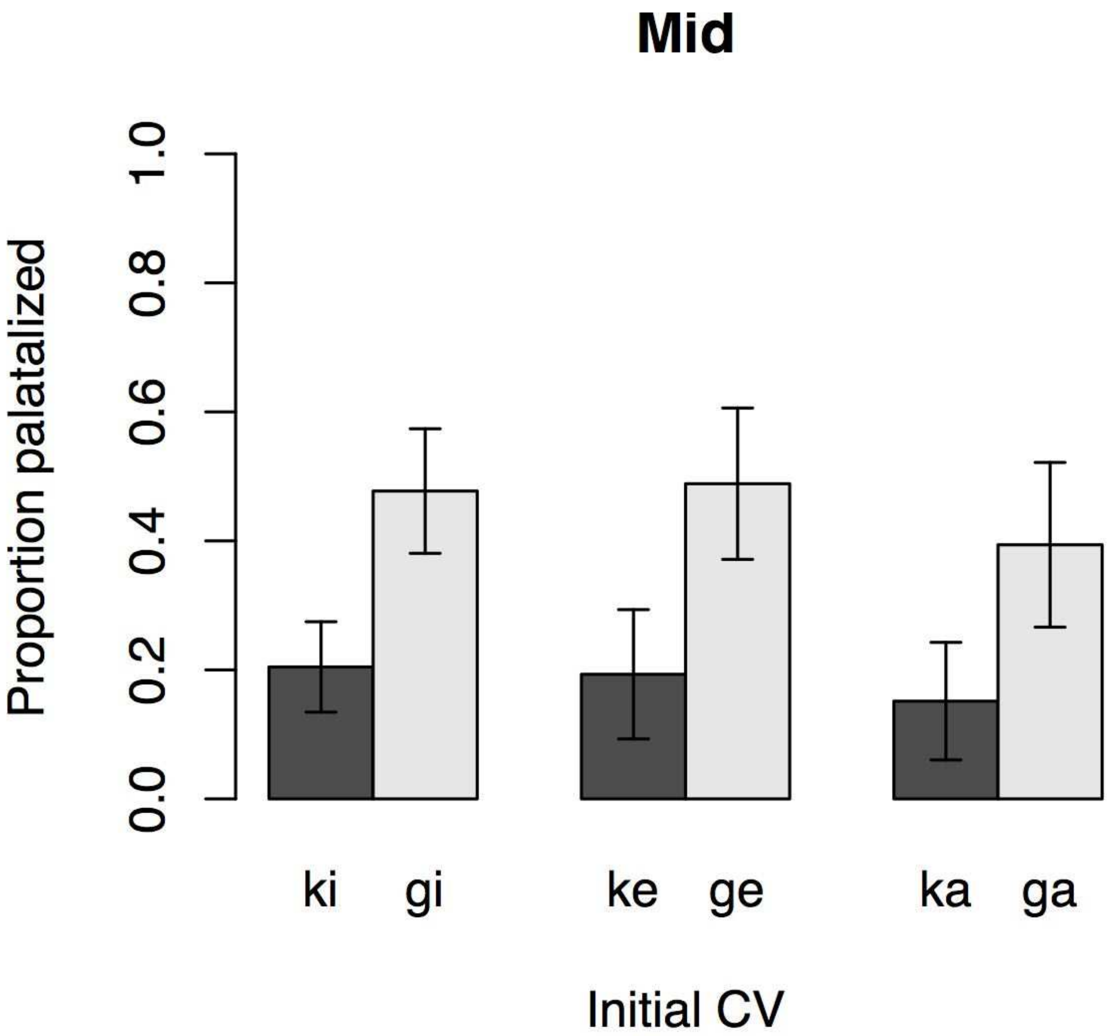
- The relevant statistic is the interaction between condition (High vs Mid) and vowel environment (exposure vs novel)

Results of Exp 1: High (N = 11)



i vs e,
$t(10) = 3.0$,
$p < .05$

Results of Exp 1: Mid (N = 11)



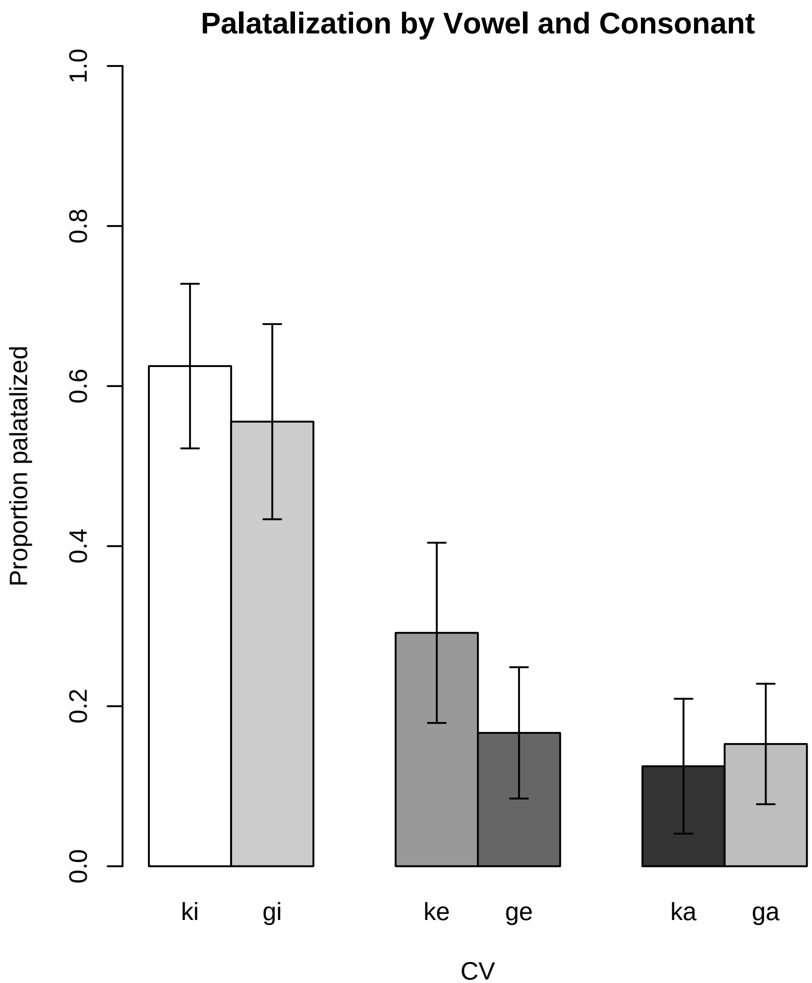
i vs e,

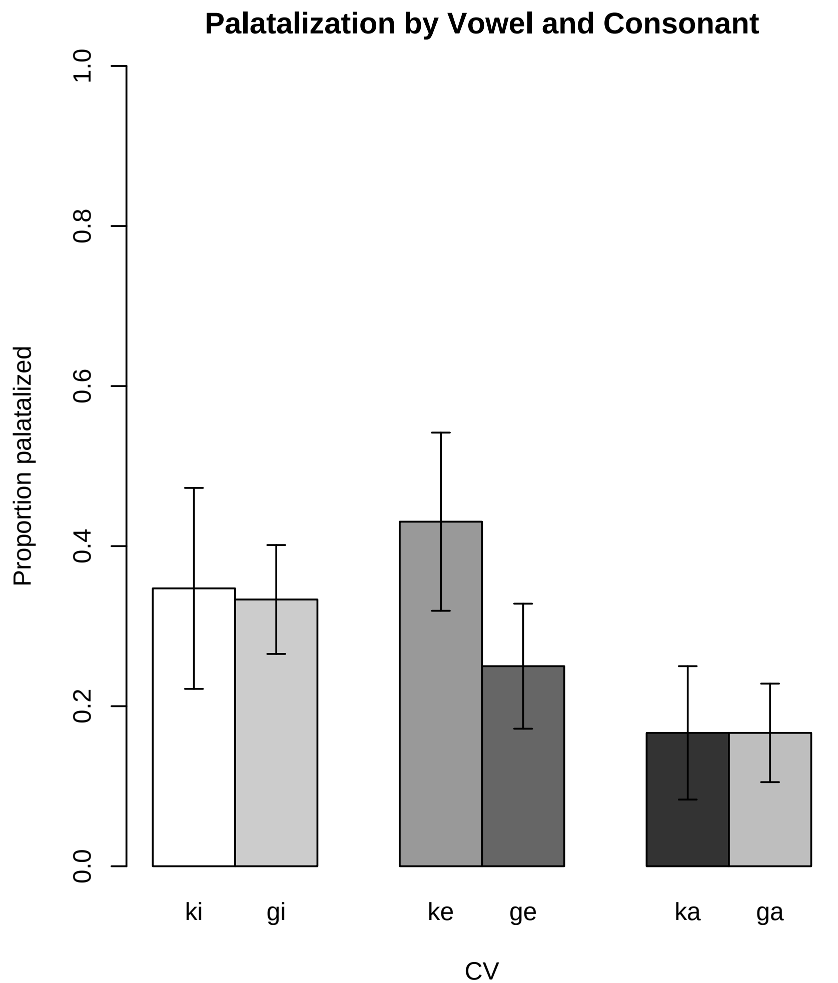
$t(10) = 0$

Results of Experiment 1

- Repeated-measures ANOVA (High vs Mid $\times$ Voiceless vs Voiced $\times$ Exposure vs Novel)
- **Greater generalization in the Mid condition than in the High condition**
  Condition $\times$ Context: $F(1, 20) = 8.3, p < .01$
- Main effect of consonant voicing (k vs g): $F(1, 20) = 8.0, p < .05$
  Why? Practice items (n=2) contained voiced g

Experiment 1A: High vs Mid





Experiment 2: Voiceless vs Voiced

- Voiceless exposure

kimə . . . t͡ʃimə $^{\times 4}$	kenə . . . t͡ʃenə $^{\times 4}$	kapə . . . kapə $^{\times 3}$
pilə . . . pilə	pebə . . . pebə	parə . . . parə

- Voiced exposure

gimə . . . d͡ʒimə $^{\times 4}$	gerə . . . d͡ʒerə $^{\times 4}$	gapə . . . gapə $^{\times 3}$
pilə . . . pilə	pebə . . . pebə	parə . . . parə

- Testing (both conditions; also included fillers)

kimə . . . ? $^{\times 8}$	kenə . . . ? $^{\times 8}$	kapə . . . ? $^{\times 3}$
gimə . . . ? $^{\times 8}$	gerə . . . ? $^{\times 8}$	gapə . . . ? $^{\times 3}$

Qualitative predictions

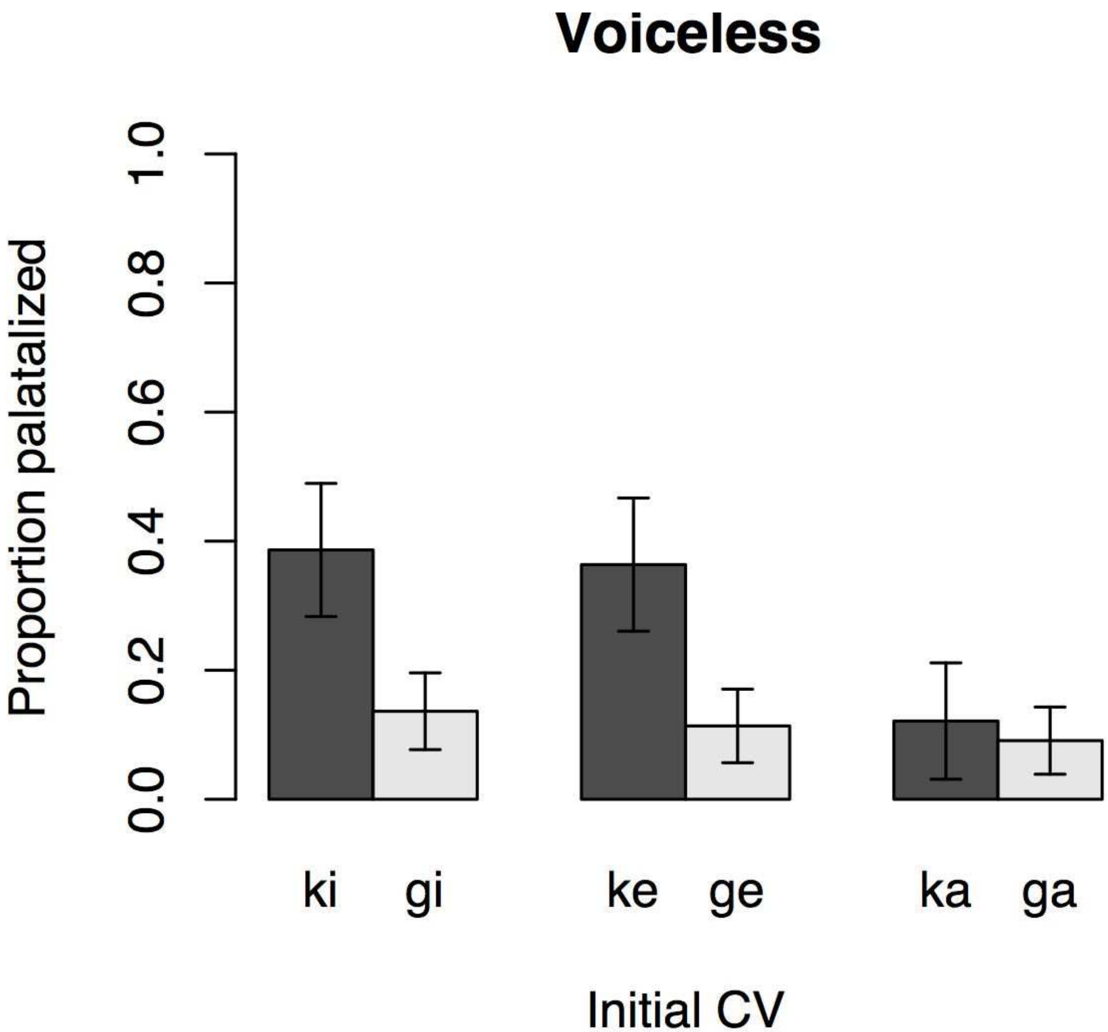
- Guion 1996, 1998 found greater similarity for **voiceless** k and t͡ʃ than for **voiced** g and d͡ʒ

+ participants in the Voiced group (g $\longrightarrow$ d͡ʒ__i/e) should generalize to the novel consonant (k)

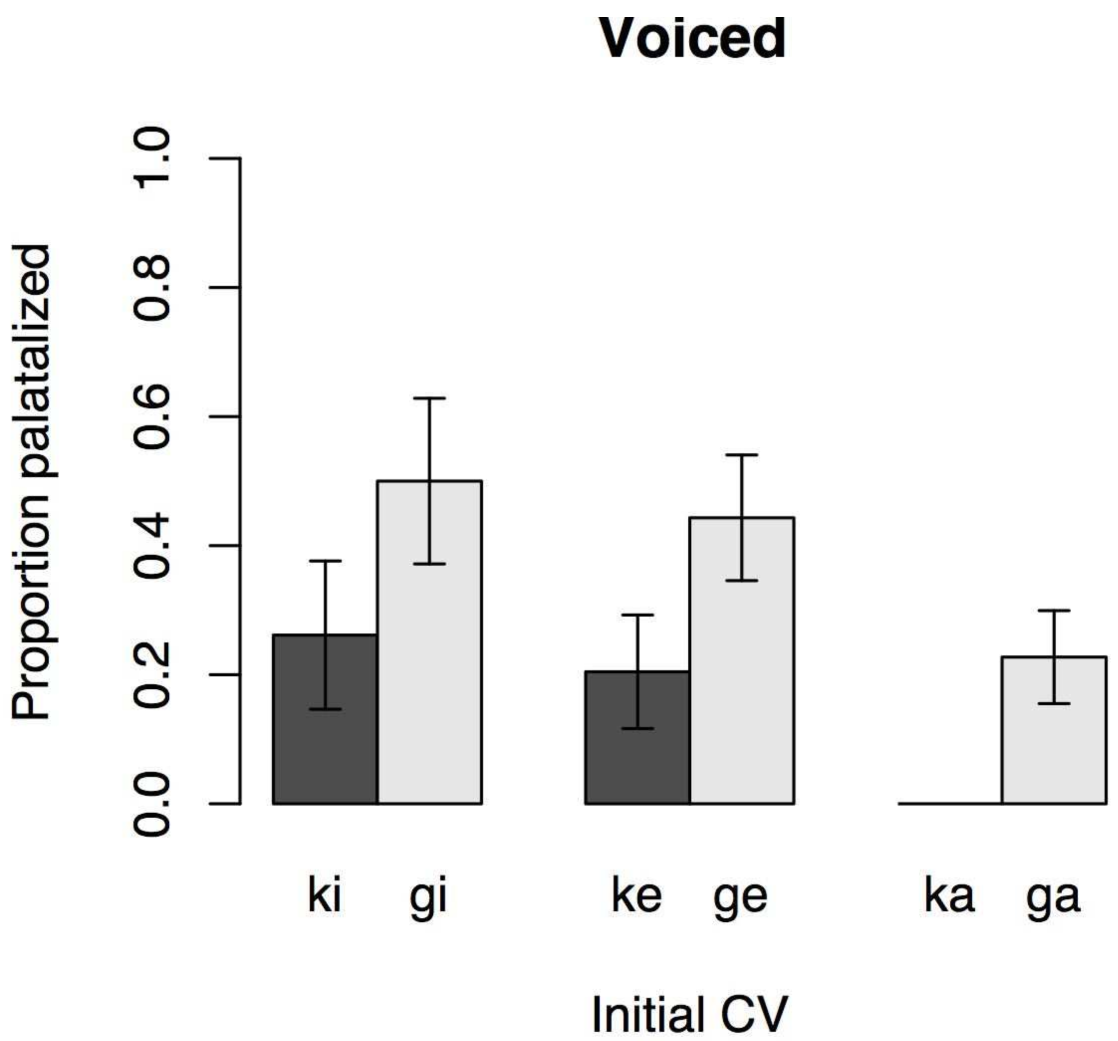
– but do not expect generalization to the same degree in the Voiceless group (k $\longrightarrow$ t͡ʃ__i/e)

- As before, the relevant statistic tests for an interaction: condition (Voiceless vs Voiced) and focus consonant (exposure vs novel)

Results of Exp 2: Voiceless (N = 11)



Results of Exp 2: Voiced (N = 11)



Results of Experiment 2

- Repeated-measures ANOVA (Voiceless vs Voiced $\times$ High vs Mid $\times$ Exposure vs Novel)
- **Non-significant difference between the rates of generalization in the conditions**
  Condition $\times$ Focus: $F(1, 20) < 1$
- All other main effects and interactions (except main effect of focus) n.s.

Summary of experiments

- Exp 1. Asymmetric generalization on the **context** is consistent with Steriade's law:
  - Δ_P(k,t͡ʃ /__ i) < Δ_P(k,t͡ʃ /__ e)

  $+$ k $\rightarrow$ t͡ʃ/__ e generalized to __i (and __ɑ!)

  $-$ k $\rightarrow$ t͡ʃ/__ i not generalized to __e

- These results are not explained by the "error" mechanisms postulated by evolutionary phon

- Exp 2. Lack of generalization on the **focus** suggests an architectural limit on the law

Modeling substantive bias

Overview

- 1. Quantifying perceptual similarity with the GCM (Nosofsky 1986, et seq)
- 2. Integrating substantive bias into the maximum entropy (maxent) formalism
- 3. Comparing predictions of the biased and unbiased with the experimental results

1. Generalized context model (GCM)

- Stimulus properties (Nosofsky 1986)

$$d_{ij} = c[\sum_{m=1}^{M} w_m |x_{im} - x_{jm}|^r]^{1/r} \qquad \sum w_m = 1$$

$$c > 0$$

- Perceptual similarity (Shepard 1957, 1987)

$\eta_{ij} = \exp(-d_{ij})$ <— we solve for this

- Luce choice rule (Luce 1962)

$$P(\text{resp} = x_j | \text{ stim} = x_i) = \frac{b_j \eta_{ij}}{\sum_{k=1}^{n} b_k \eta_{ik}} \qquad \sum b_k = 1$$

1. Applying the GCM

- How perceptually similar are velar stops and palatoalveolar affricates in vowel contexts?
- Stimulus dimensions
  peak spectral frequencies (Guion / ours)
  dummy-coded vowel quality and voicing
- Confusion matrix (Guion)
  [contains data for i and ɑ, but not e]
- Maximum-likelihood fit of scale (c), attention weights ($\{w_m\}$), and response biases ($\{b_k\}$)

1. Response bias $*$ similarity values

Inverse of the perceptual ‘cost’ of changing a velar stop to a palatoalveolar affricate

- $\frac{\text{Response bias(t͡ʃV)} * \text{similarity(kV,t͡ʃV)}}{}$

ki → t͡ʃi	ke → t͡ʃe	kɑ → t͡ʃɑ
9.23^{-1}	$\mathit{12.68}^{-1}$	88.72^{-1}

- $\frac{\text{Response bias(d͡ʒV)} * \text{similarity(gV,d͡ʒV)}}{}$

gi → d͡ʒi	ge → d͡ʒe	gɑ → d͡ʒɑ
21.13^{-1}	$\mathit{40.60}^{-1}$	126.93^{-1}

- Cost for e context estimated with b(Xe) = b(Xi)

2. Maxent model: General

- Type of log-linear model in which entropy is maximized subject to (expected constraint violation) = (observed constraint violation)

- Closely related to random fields (Della Pietra et al. 1986, Lafferty et al. 2001, Johnson et al.) and Harmony theory (Smolensky 1986)

- OT as a limiting case: higher-ranked constraints have infinitely stronger weights (sometimes realizable with exp weighting)

2. Maxent model: Constraints

- Faithfulness (output-to-output)
  F(k) violated by the change k $\longrightarrow$ t͡ʃ
  F(g) violated by the change g $\longrightarrow$ d͡ʒ

- Markedness
  M(C / V) violated by C (k or g) followed by vowel with features V ($[\pm high, \pm low]$)

Features: i $[+high, -low]$, e $[-high, -low]$, ɑ $[-high, +low]$

2. Maxent model: output probability

- Assume that experimental design restricts candidate set to two outputs

$\mathbf{y}^{pal}$	palatalized	ex. kimə . . . t͡ʃimə
$\mathbf{y}^{faith}$	faithful	ex. kimə . . . kimə

- Probability of palatalizing stimulus $\mathbf{x}$

$$P(\mathbf{y}^{pal}|\mathbf{x}) = \frac{H(\mathbf{x},\mathbf{y}^{pal})}{H(\mathbf{x},\mathbf{y}^{pal}) + H(\mathbf{x},\mathbf{y}^{faith})}$$

where $H(\mathbf{y}|\mathbf{x})$ is the **harmony** of $\mathbf{y}$ given $\mathbf{x}$ . . .

2. Maxent model: harmony

- Harmony is an exp function of the sum of weighted (λ_k) constraint violations (f_k)

$$H(\mathbf{x},\mathbf{y}) = \exp(-\sum_{i=1}^{K} \lambda_k f_k(\mathbf{x},\mathbf{y}))$$

- $H(\mathbf{x},\mathbf{y}^{pal})$ = $\begin{cases} \lambda_{\mathrm{F(k)}}, \text{ if } \mathbf{x} \text{ begins with k} \\ \lambda_{\mathrm{F(g)}}, \text{ if } \mathbf{x} \text{ begins with g} \end{cases}$

- $H(\mathbf{x},\mathbf{y}^{faith})$ = $\sum\{\lambda_{\mathrm{M(C/V)}} |$ $\mathbf{x}$ begins with CV $\}$

2. Maxent model: training

- Model trained on the same $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ pairs as the experiment participants (for D iterations)

- Objective (Lafferty et al. 2001, McCallum 2003)

  $-D \sum_{i=1}^{N} \log P_\Lambda(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$ (pseudo-likelihood)

  $+\sum_{k=1}^{K} \frac{(\lambda_k - \mu_k)^2}{2\sigma_k^2}$ (Gaussian prior/regularizer)

- Convex optimization problem $\Rightarrow$ global min

2. Maxent model: substantive prior

- Given similarities calculated in (1), how can we incorporate Steriade's law into the prior?

  $\sum_{k=1}^{K} \frac{(\lambda_k - \mu_k)^2}{2\sigma_k^2}$ (from previous slide)

- Assign mean ranking value $\mu = 0$ to all M(C/V) constraints — consistent with L1

- **Penalize deviation from μ in proportion to the cost of changes that satisfy M(C/V)**

  $\sigma_k \leftarrow \min\{$ biased-sim(velar$_k \rightarrow$ pal $|$ V$_k$)$\}$

2. Maxent model: substantive prior

Example calculations

M(k / $[+high, -low]$) — *ki

$\sigma = \min\{\ b(\text{t͡ʃi})\eta(\text{ki,t͡ʃi})\ \} = \min\{9.23^{-1}\} = 9.23^{-1}$

M(k / $[-low]$) — *ki, ke

$\sigma = \min\{\ b(\text{t͡ʃi})\eta(\text{ki,t͡ʃi}), b(\text{t͡ʃe})\eta(\text{ke,t͡ʃe})\} =$
$\min\{9.23^{-1}, 12.68^{-1}\} = 12.68^{-1}$

$\Rightarrow$ Penalty for deviating from mean (0) weight is approximately 2 times greater for M(k / $[-low]$)

2. Understanding the substantive bias

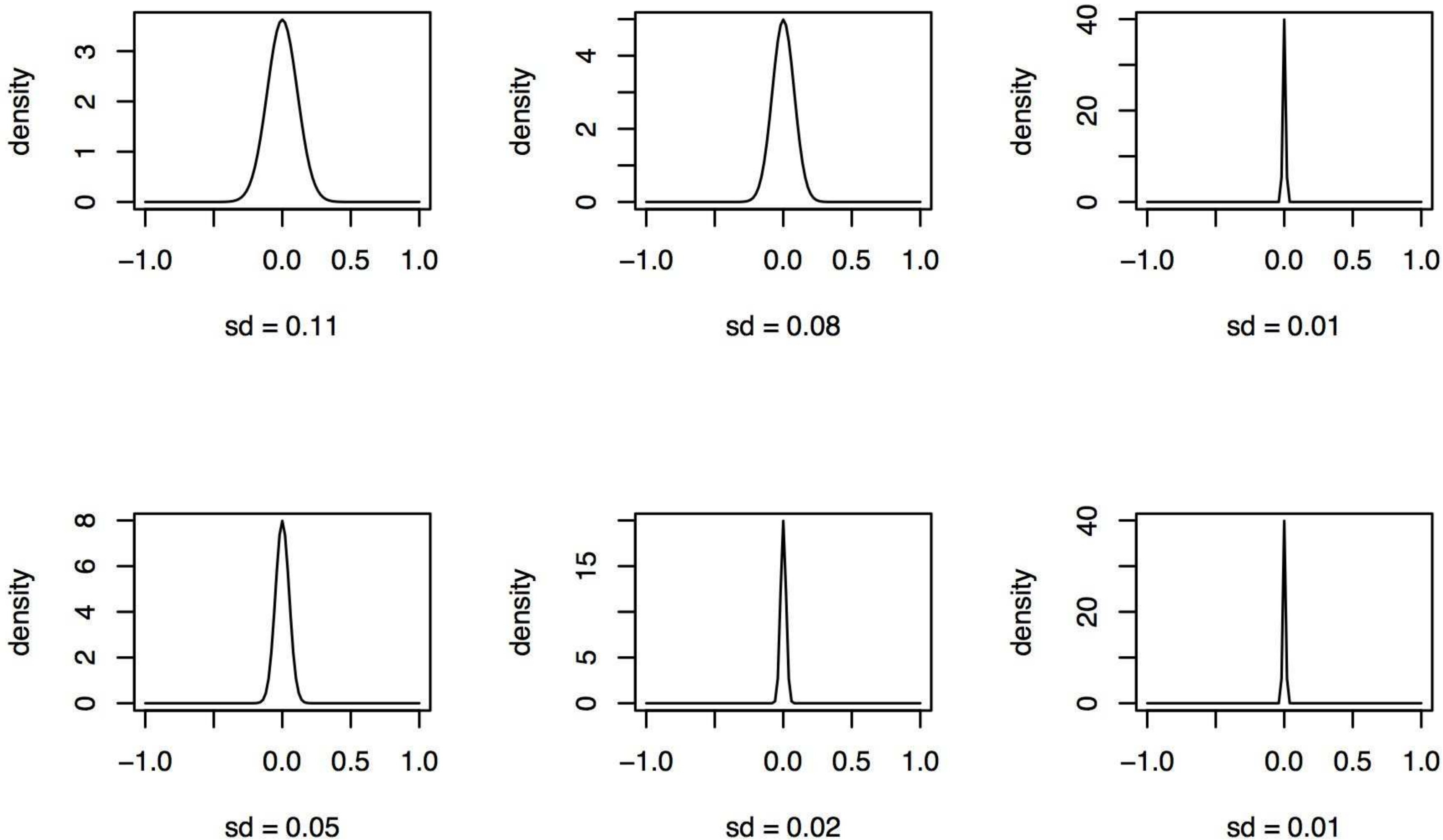
- All Markedness constraints considered here have a mean/default ranking values of 0

  Set output-to-output Faithfulness much higher (10.0)

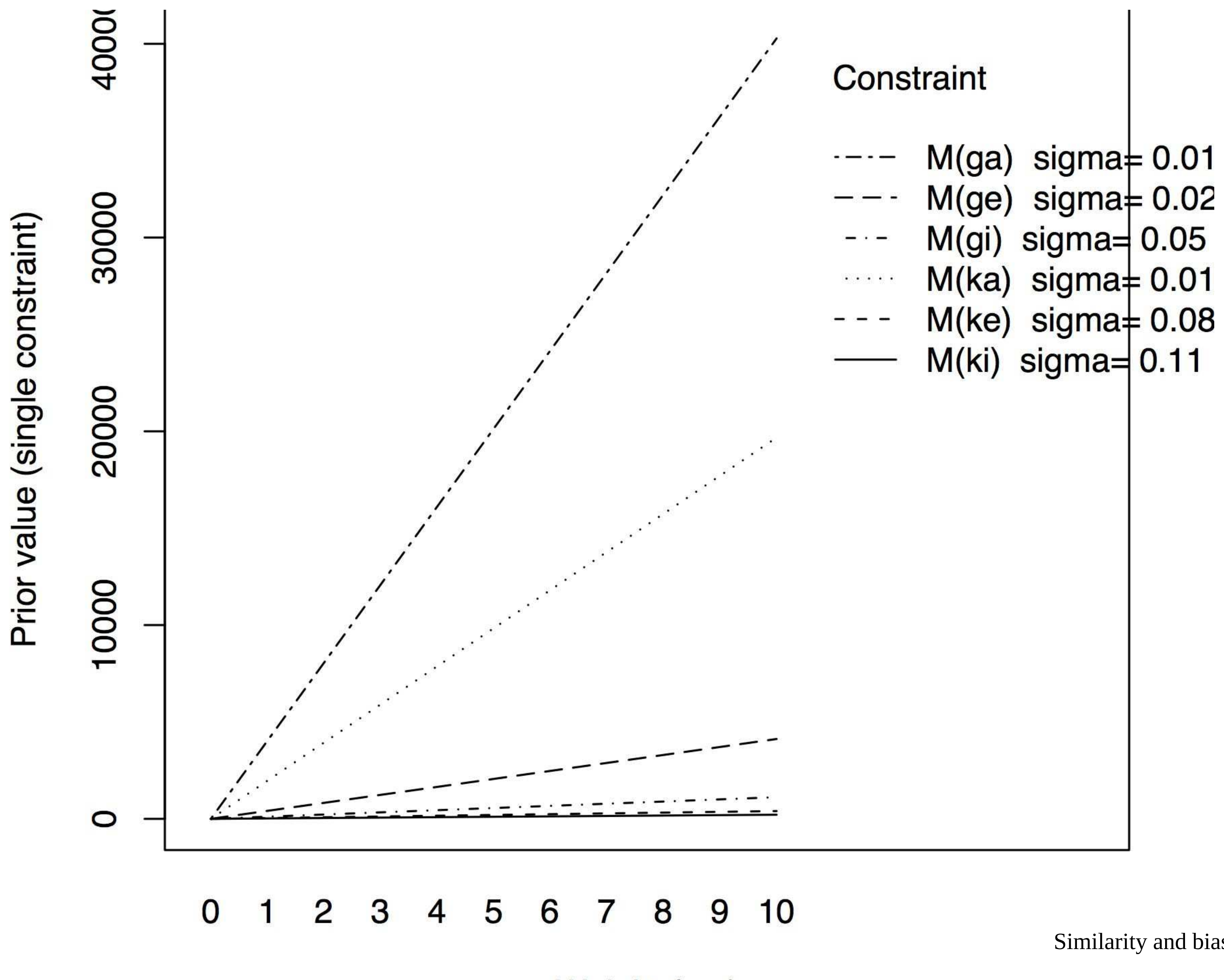
- Constraints with 0 weight have no effect on the output probabilities, hence the default output is faithful (no palatalization)

- Markedness constraints that compel more perceptually costly palatalization receive greater penalties for deviating from 0 weight

2. Prior density by σ



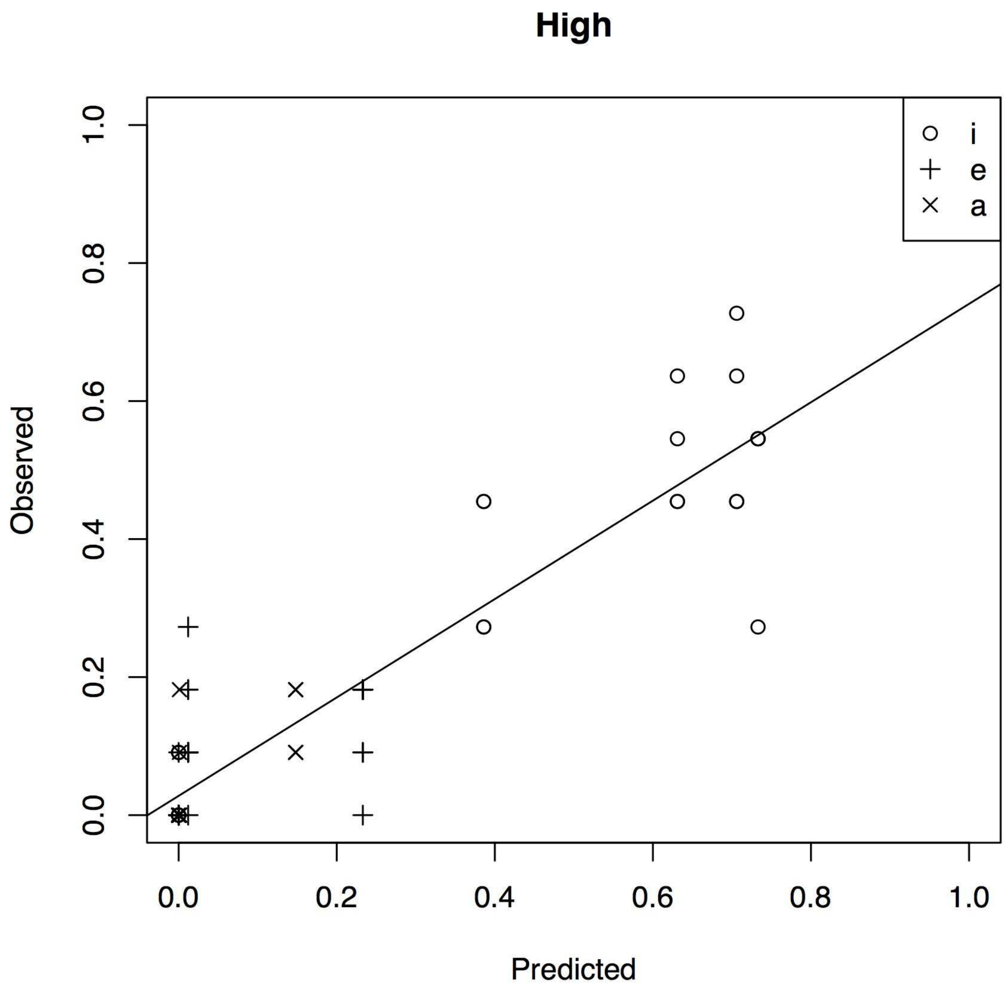
2. Prior penalty by σ and weight (λ)



2. Unbiased alternative

- In the substantively-biased model, all Faith constraints have a relatively large σ $(.01)$
- An unbiased version of the model is easy to construct: simply assign the same $\sigma = .01$ to all Markedness constraints as well
- The next section compares the predictions of the biased and unbiased models with the results of Experiment 1 and 2

3. Experiment 1: High



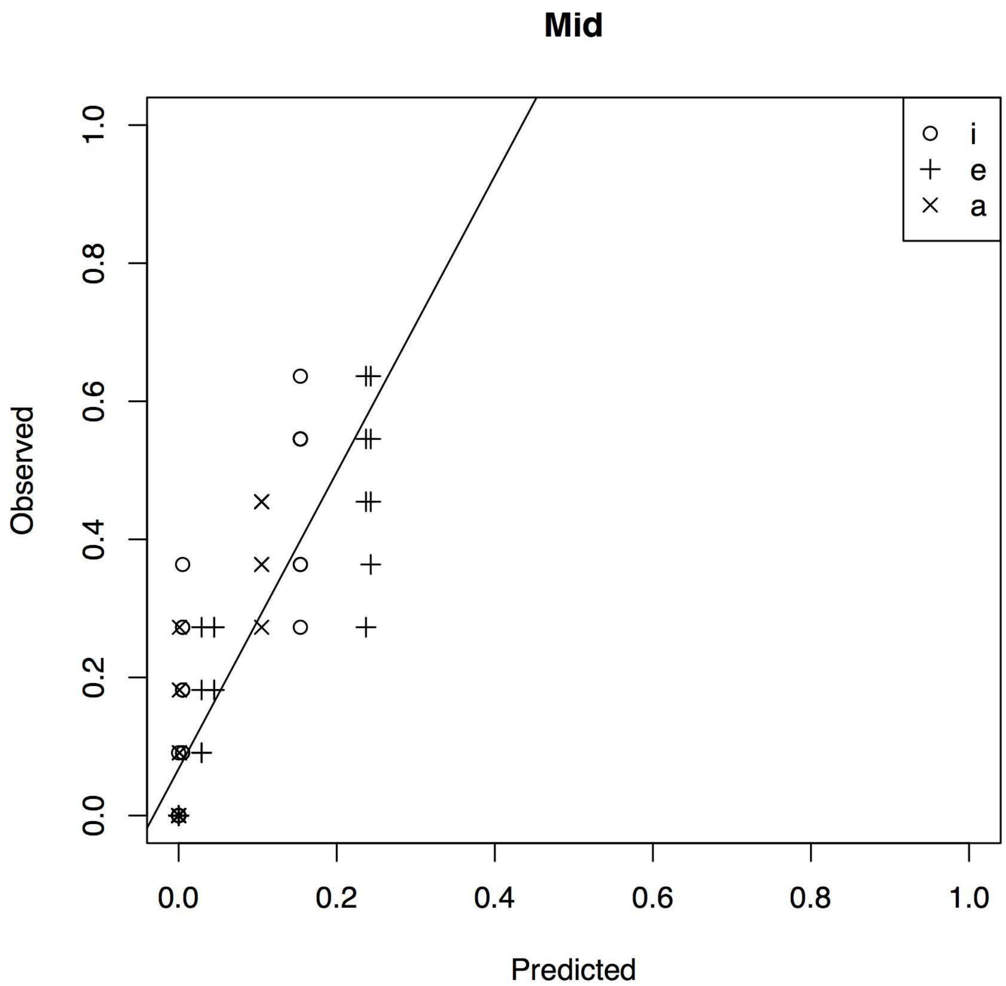
biased

$r = .910$ (82.8%)

cf. unbiased

$r = .913$ (83.4%)

3. Experiment 1: Mid



biased

$r = .859\ (73.8\%)$

cf. unbiased

$r = .550\ (30.3\%)$

3. Bias and asymmetric generalization

- High condition. The prior σ values of M(k/$[+hi, -lo]$) and F(k) are approx. equal, therefore the constraints "meet in the middle"

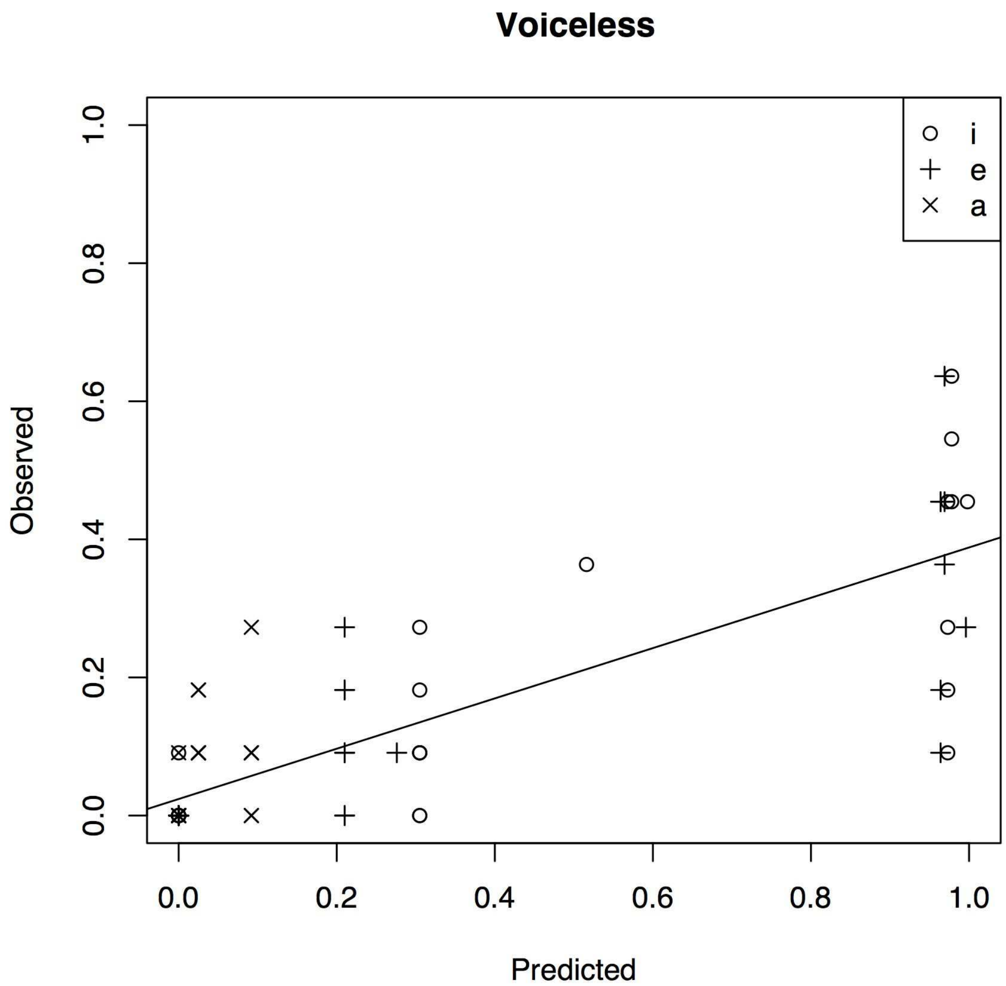
  M$'$ $<$ M(k/$[+hi, -lo]$) $\approx$ F(k)

- Mid condition. But the prior σ value of M(k/$[-hi, -lo]$) is much smaller than that of F(k), so the latter gets "dragged down"

  M$'$ $\approx$ M(k/$[-hi, -lo]$) $\approx$ F(k)

- Cf. unbiased model never generalizes

3. Experiment 2: Voiceless



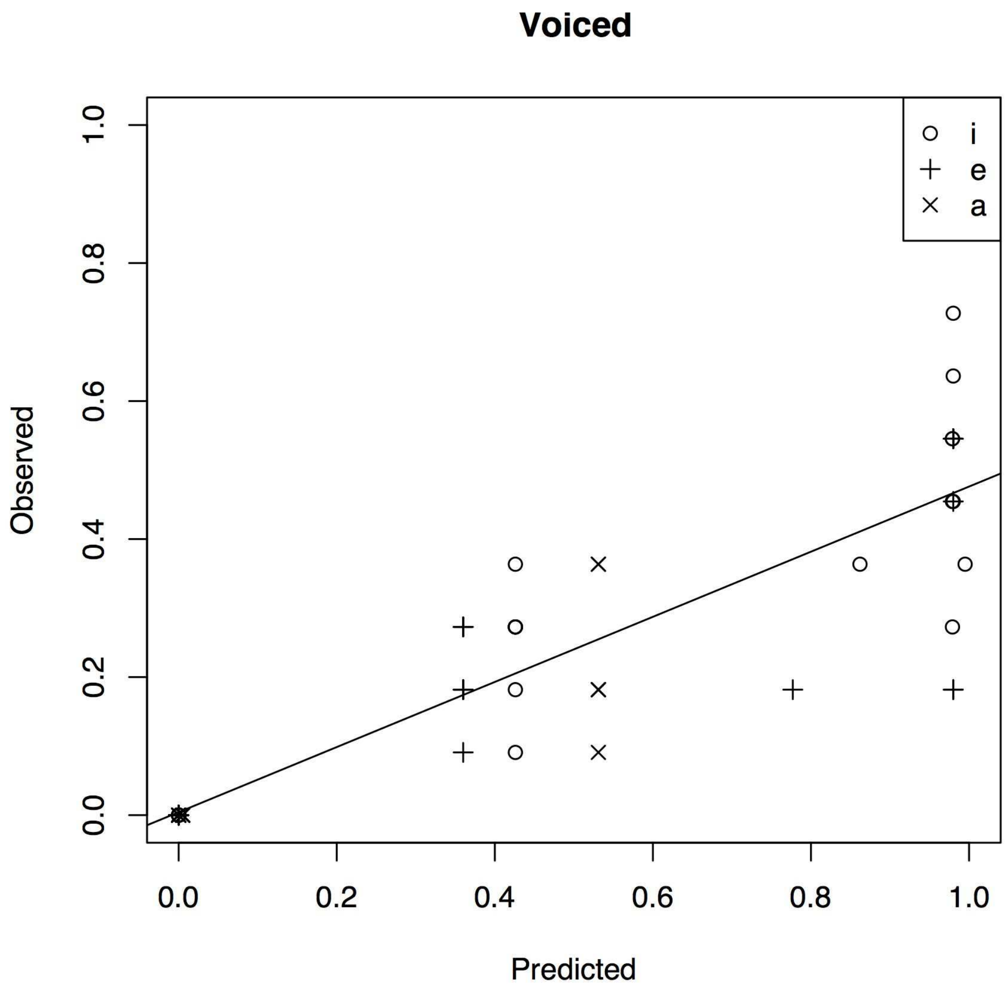
biased

$r = .807\ (65.1\%)$

cf. unbiased

$r = .811\ (65.8\%)$

3. Experiment 2: Voiced



biased

$r = .920\ (84.6\%)$

cf. unbiased

$r = .875\ (76.6\%)$

Discussion and conclusions

Evidence for substantive bias

- Substantively-biased model, but not unbiased model, accounts for the asymmetric generalization in Exp 1
- Difference in the Mid condition — the only one showing significant generalization — is stark (approx. 30% of variance)
- Modeling of Exp 2 results shows how behavior depends on prior *and* constraints

Discussion: hard vs soft bias

- Substantively-biased phonology does not impose hard restrictions on phono patterns
- Given sufficient high-quality input, even g → d͡ʒ __ ɑ is learnable (“overwhelming the prior”)
- This mitigates the empirical problems — attested random rules — that plague more rigid phonetically based proposals

Discussion: types of generalization

- Why do we observe generalization on the **context**, but not on the **focus**?
- Similar results (for k g) found in speech-error experiments of Goldrick 2004
- Plausibly due to an architectural feature motivated, on average, by economy

  changes can be identified (listed) with few features

  must generalize to compactly describe contexts

Discussion: relationship to OT

- Empirical difference between (stochastic) OT and maxent w.r.t. **harmonic bounding**
- These experiments sidestep the issue: candidates violate constraints at most once
- Assuming harmonically bounded candidates always lose, can we formulate stochastic OT learning as a convex optimization problem?

Summary

- What is the relationship between phonetics and phonology?
  Knowledge of phonetics (here, perception) biases the symbolic system in favor of states that are functionally motivated

- Substantive bias can be formalized with standard methods from psych and ling

- The nature of how learners **generalize** from (possibly quite limited) exposure is the central theoretical issue of phonology