LSA 511: Computational Models of Sound Change

James Kirby & Morgan Sonderegger

11 Jul 2013

Phonologization

- *Phonologization*: how phonetic variation → phonological
- Two traditions:
  - Phonetic (e.g. Ohala 1981): focus on individuals; phonetic variables; continuous parameter spaces
  - Socio-historical (e.g. Weinrich et al. 1968, Kroch 1989, Labov 2001): focus on speech communities; grammatical change
- Today: continuous parameters in populations.

Phonologization through coarticulation

(Some? all?) systematic phonological variation originates in phononologization of phonetic detail, especially coarticulation (Baudouin 1895; Öhman, 1966; Ohala, 1981; Blevins, 2004....)

WGmc	*Pre-OHG*	*OHG (NHD)*
*gasti	gesti	gest (*Gäste)*
*lambir	lembir	lemb (*Lämme)*
*fasti	festi	fest (*fest)*

Primary umlaut in West Germanic (after Iverson and Salmons, 2006).

For Germanicists (from Iverson & Salmons 2003)

(1) Primary umlaut, OHG
- a. *gast* ∼ *gesti* 'guest, guests'
- b. *lamb* ∼ *lembir* 'lamb, lambs'
- c. *fasto* ∼ *festi* 'solid/fast', adv. and adj.

(2) Blocking of primary umlaut, OHG (but 2ndary MHG umlaut)
- a. *maht* ∼ *mahti* 'power, powers' (also dialect. *mehti*)
- b. *haltan* ∼ *haltis* 'to hold, you hold' (also dialect. *heltis*)
- c. *starch* ∼ *starchio* 'strong, stronger' (also *sterchio*)

(3) General fronting of all back vowels before /i,j/
- a. OHG *gruoni*, MGer *grün* 'green'
- b. OHG *skoni*, MGer *schön* 'beautiful'

The listener as a source of sound change (Ohala, 1994)

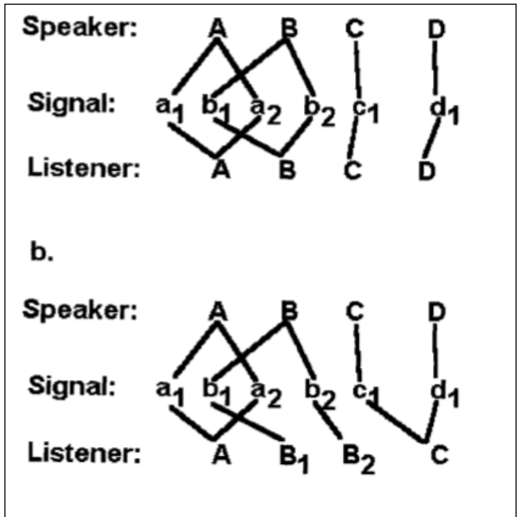
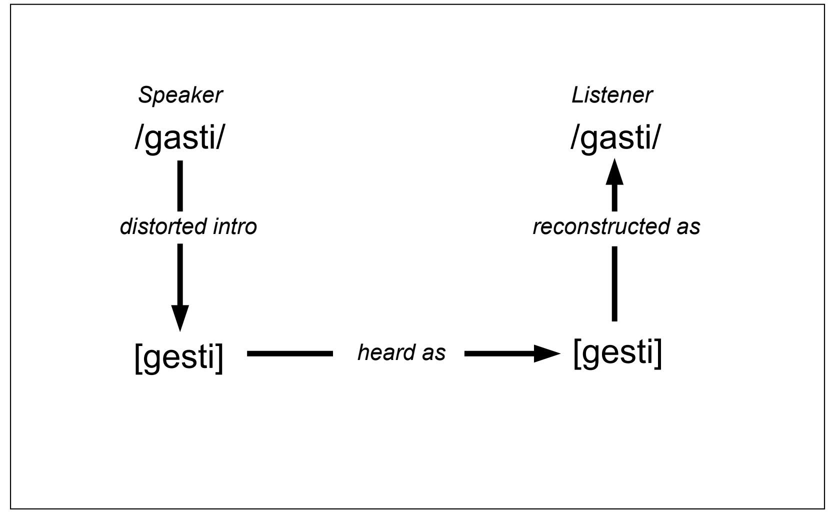
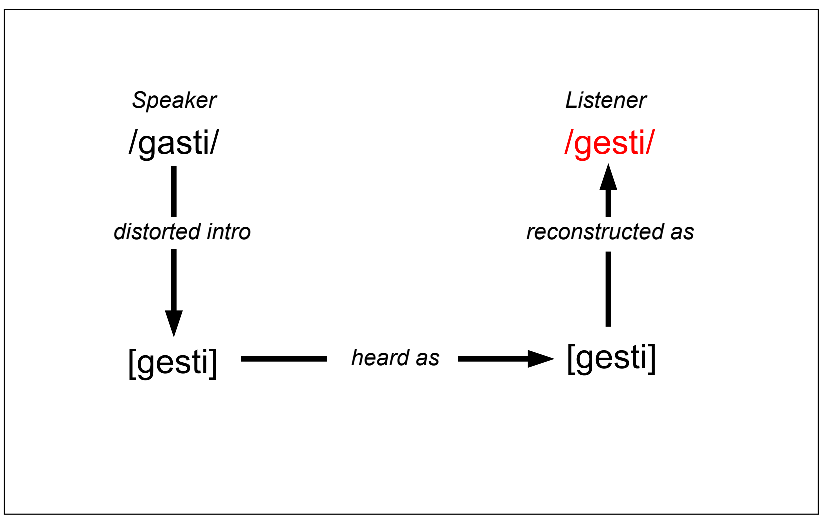


Fig. 1. How the listener might parse the speech signal.

The listener as a source of sound change (Ohala, 1994)



The listener as a source of sound change (Ohala, 1994)



From individuals to populations

- 'I assume without further argument that the initiation of such sound changes is accomplished by the phonetic mechanism just described; their spread, however, is done by social means, e.g., borrowing, imitation, etc.' (Ohala 1981: 184)
- Is such a mechanism plausible in social (as opposed to iterated) learning scenarios (Niyogi & Berwick, 2009)?

Phonologization at the population level

- The mere *presence* of a potential trigger does not imply that phonetic change is inevitable (Kiparsky, 1995)
- Default is stability, not change! (Weinrich et al, 1968)
- Change in a single individual is neither necessary nor sufficient for diachronic change
- If 'initiation' (actuation) is at the level of the individual, under what conditions will it take hold in a population?

Learning: continuous vs. discrete

- Learning of phonetic category structure (VOT, F1) is effectively continuous (maybe)
- Tuesday's case: learning a continuous probability over discrete selection space
- Here: both data and parameter being learned are continuous

Learning, in general

- In $\mathcal{G}_{t+1}$, each learner is presented with n examples drawn from teachers in $\mathcal{G}_t$ chosen by some sampling procedure $\mathcal{S}$.
- Learner applies some learning algorithm $\mathcal{A}$.
- Assuming $\mathcal{S}$ and $\mathcal{A}$ are the same for all learners in $\mathcal{G}_{t+1}$, this implies the following evolution equation for π_t:

$$(\pi_{t+1}) = f_{S,\mathcal{A}}(\pi_t \,|\, \text{constants})$$

Learning: continuous vs. discrete

- This holds *in general* but the form of f can be rather different
- Discrete pmf (from Tues):

$$\Pr(X = k) = \binom{N}{k} (\pi_t)^k (1 - \pi_t)^{N-k}$$

- Continuous pdf:

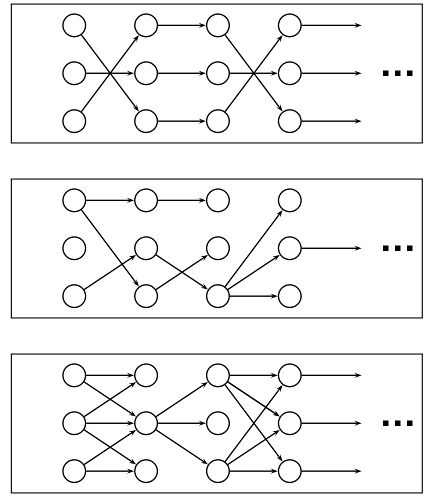
$$\Pr(a \leq X \leq b) = \int_a^b f(x)\, dx$$

- where f is Gaussian, Poisson, or something

Stability and change

- Goal: explore effects of different assumptions about bias and population structure on the evolution of a continuous phonetic parameter of the sort presupposed in the previous section
- Under what assumptions can we predict
  1. Stability of limited coarticulation in the population.
  2. Stability of full coarticulation in the population (e.g. umlaut).
  3. Change from stable limited to full coarticulation.

Properties of populations



(a) Parallel diffusion chains (classic IL)

(b) Single-teacher learning

(c) Multiple-teacher learning

Framework

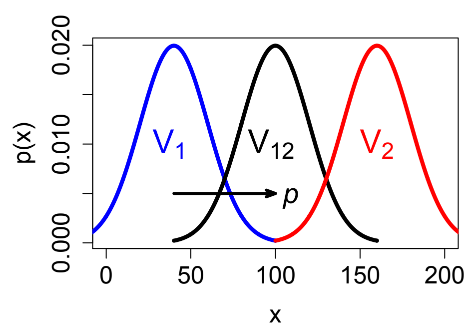
Some assumptions:

- speech sounds have been organised into discrete segments
- phonetic realisation of segments is subject to coarticulation
- learners have access to the complete segmental inventory

Framework

Lexicon $\Sigma = \{\mathrm{V}_1, \mathrm{V}_2, \mathrm{V}_{12}\}$,

where V_{12} represents V_1 in the coarticulation-inducing context of V_2 (e.g., [a_i], as opposed to plain [a] or [i])



- Task: infer offset parameter p from sample $\bar{y}$, where $P(y_i) \sim N(\mu_a - p, \sigma_a^2)$
- State of population at t wholly characterized by distribution of $p \sim \pi_t(p)$

Framework

Even more assumptions:

- for a given learner, categories are distributed

$$\mathsf{V}_1 \sim N(\mu_a, \sigma_a^2), \mathsf{V}_2 \sim N(\mu_b, \sigma_b^2), \mathsf{V}_{12} \sim N(\mu_a - p, \sigma_a^2)$$

- learners are divided into discrete generations $\mathcal{G}_t$ of size M
- assume M is infinite, so evolution of the population is not a stochastic process

Models

MaxLik

no prior
($\mathcal{A}_{\text{naive}}$)

MAP/EV

simple prior
($\mathcal{A}_{\text{simple}}$)

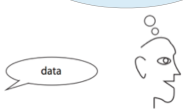
Polynomial

complex prior
($\mathcal{A}_{\text{complex}}$)

Models

MaxLik

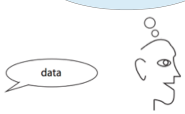
no prior
($\mathcal{A}_{\text{naive}}$)



$$\hat{p} = \mu_a - \bar{y}$$

MAP/EV

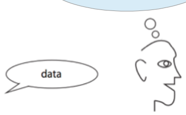
simple prior
($\mathcal{A}_{\text{simple}}$)



$$\hat{p} = \frac{(\mu_a - \bar{y})}{1 + \sigma_a^2 / \; n\tau^2}$$

Polynomial

complex prior
($\mathcal{A}_{\text{complex}}$)



$$\hat{p} = \text{intractable}$$

$\mathcal{A}_{\text{naive}}$

- Given a training sample $\vec{y} = (y_1, \ldots, y_n)$, the learner's maximum-likelihood estimate of p is

$$\hat{p} = \mu_a - \bar{y}$$

- The (noisy) distribution of values they could learn is then

$$P(\hat{p} \,|\, p_{\text{parent}}) = \mathcal{N}(p_{\text{parent}}, \sigma_a^2 / n)$$

- Let's just consider the evolution of mean and variance...

Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}$

- Expected value of $\hat{p}$ is

$$\begin{aligned} E[\hat{p}] &= \int \pi_{t+1}(\hat{p})\,\hat{p}\,d\hat{p} \\ &= E[p] \end{aligned}$$

- The variance of $\hat{p}$ is

$$\begin{aligned} \text{Var}(\hat{p}) &= E[(\hat{p} - E[\hat{p}])^2] \\ &= E[\hat{p}^2] - E[\hat{p}]^2 \\ &= \sigma_a^2/n + \text{Var}(p) \end{aligned}$$

Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}$

- Expected value of $\hat{p}$ is

$$\begin{aligned} E[\hat{p}] &= \int \pi_{t+1}(\hat{p})\,\hat{p}\,d\hat{p} \\ &= E[p] \end{aligned}$$

- The variance of $\hat{p}$ is

$$\begin{aligned} \text{Var}(\hat{p}) &= E[(\hat{p} - E[\hat{p}])^2] \\ &= E[\hat{p}^2] - E[\hat{p}]^2 \\ &= \sigma_a^2/n + \text{Var}(p) \end{aligned}$$

Example: normally distributed p

- Suppose p_{parent} is normally distributed in generation t:

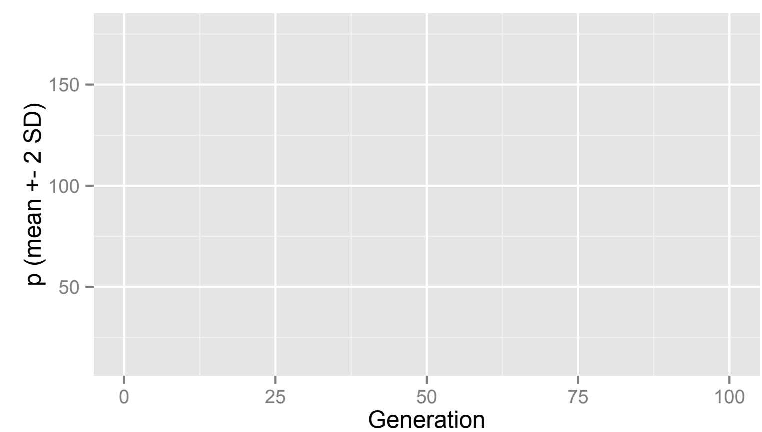
$$p_{\text{parent}} \sim \mathcal{N}(p_0, \sigma_0^2)$$

- Then $\hat{p}$ is also normally distributed:

$$\hat{p} \sim \mathcal{N}(p_0, \sigma_0^2 + \sigma_a^2/n)$$

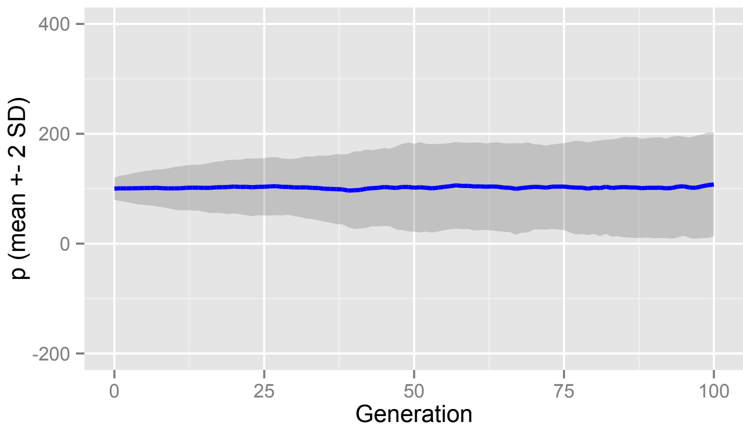
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

100 epochs, 100 examples, 1000 agents



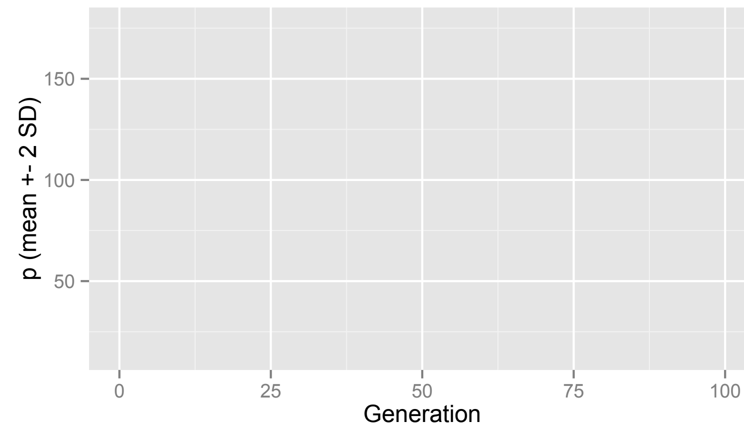
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

100 epochs, 100 examples, 1000 agents



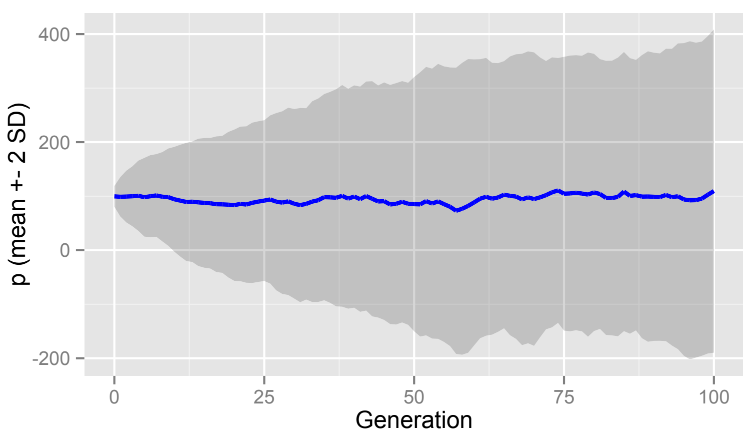
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

100 epochs, 10 examples, 1000 agents



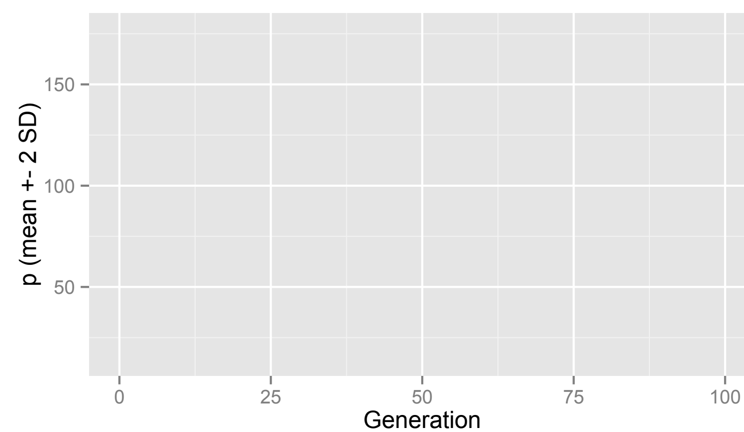
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

100 epochs, 10 examples, 1000 agents



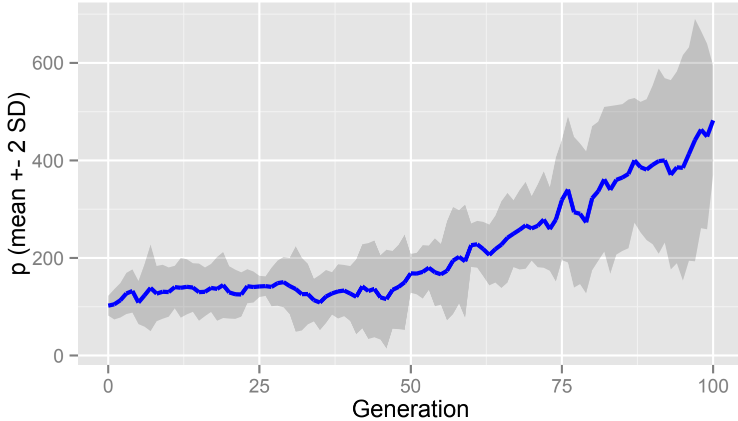
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

100 epochs, 10 examples, 10 agents



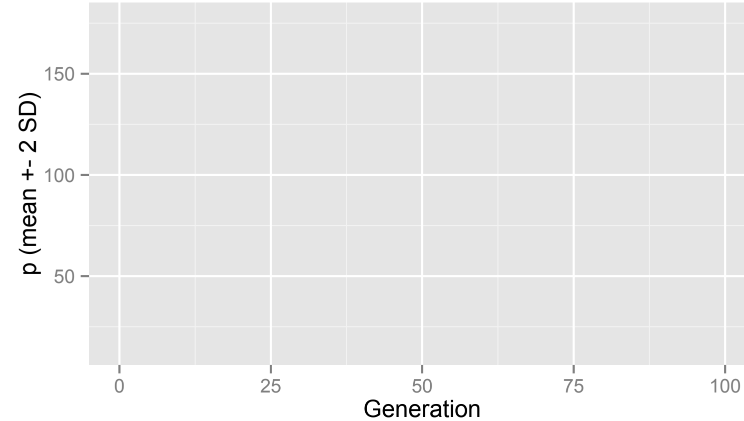
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

100 epochs, 10 examples, 10 agents



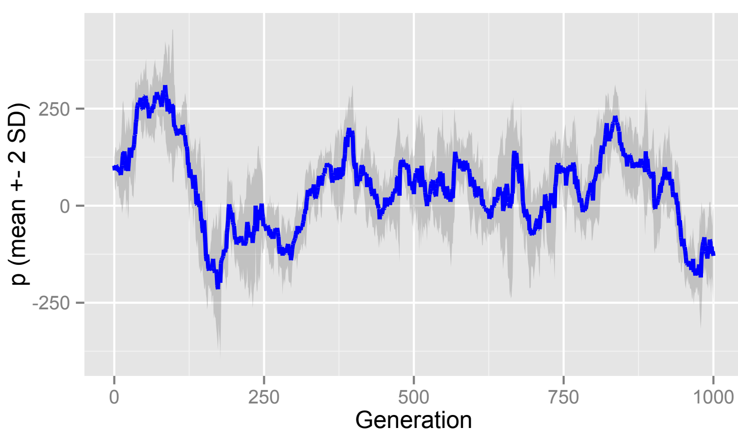
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

1000 epochs, 10 examples, 10 agents



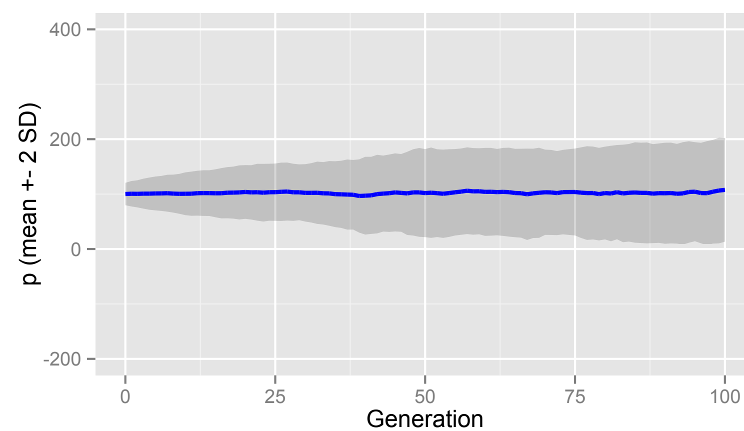
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

1000 epochs, 10 examples, 10 agents



Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{single}}, p_0 \sim N$

Holds for arbitrary distributions!



Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{multiple}}$

- Now: learners in $\mathcal{G}_{t+1}$ receive each training example from a randomly-chosen teacher in generation $\mathcal{G}_t$
- The ML estimate is still $P(\hat{p}\,|\,\vec{p}) = \mathcal{N}(\mu_a - \bar{y}, \sigma_a^2/n)$
- However, the expected value and variance of $\hat{p}$ are now

$$E(\hat{p}) = E(p_t), \quad \text{Var}(\hat{p}) = \sigma_a^2/n + \text{Var}(p_t)/n$$

- Var moves to

$$\alpha_* = \sigma_a^2/(n-1)$$

Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{multiple}}$

- Now: learners in $\mathcal{G}_{t+1}$ receive each training example from a randomly-chosen teacher in generation $\mathcal{G}_t$
- The ML estimate is still $P(\hat{p}\,|\,\vec{p}) = \mathcal{N}(\mu_a - \bar{y}, \sigma_a^2/n)$
- However, the expected value and variance of $\hat{p}$ are now

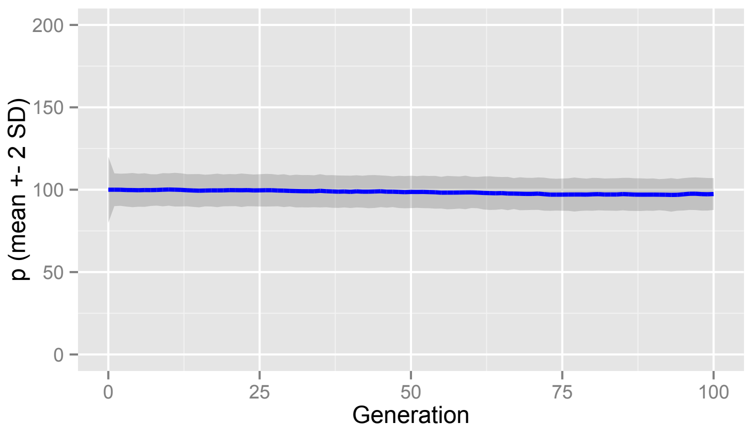
$$E(\hat{p}) = E(p_t), \quad \text{Var}(\hat{p}) = \sigma_a^2/n + \text{Var}(p_t)/n$$

- Var moves to

$$\alpha_* = \sigma_a^2/(n-1)$$

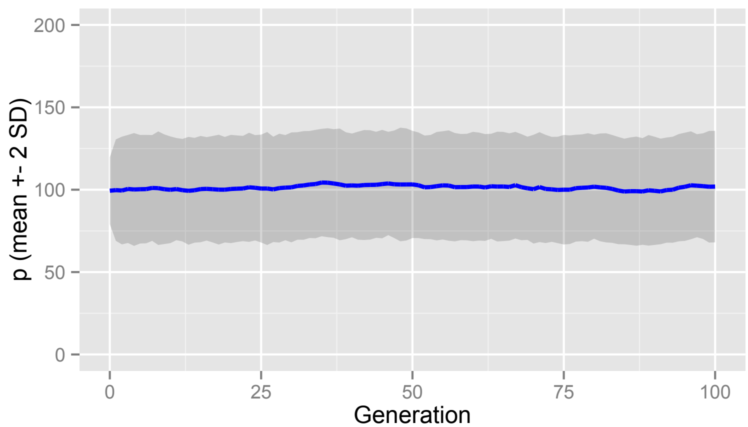
Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{multiple}}, p_0 \sim N$

100 epochs, 100 examples, 1000 agents



Evolution of π_t under $\mathcal{A}_{\text{naive}}, \mathcal{S}_{\text{multiple}}, p_0 \sim N$

100 epochs, 10 examples, 1000 agents

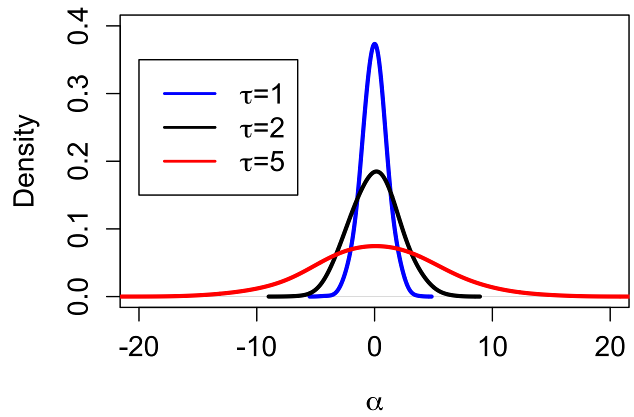


Summary: $\mathcal{A}_{\text{naive}}$

- $\mathcal{A}_{\text{naive}}$ has different dynamics under $\mathcal{S}_{\text{single}}$ and $\mathcal{S}_{\text{multiple}}$
- Under $\mathcal{S}_{\text{single}}$, variability increases w/each generation, i.e. speakers come to coarticulate more and more differently
- Under $\mathcal{S}_{\text{multiple}}$, mean of the distribution of p stays the same over time, but its variance moves towards a single value α_*.
- Both are clearly inadequate: but why do naive models fail?

$\mathcal{A}_{\text{simple}}$

- No force counteracting the noise in each learner's estimate
- Assume simple prior bias on p: $\alpha \sim \mathcal{N}(0, \tau^2)$



$\mathcal{A}_{\text{simple}}$

- How should the learner estimate $\hat{p}$?

$$\hat{p}_{\text{EV}} = \int P(p|\vec{y})p\,dp \qquad \hat{p}_{\text{MAP}} = \arg\max_p P(\vec{y}|p)P(p)$$

- Both estimates turn out to be equivalent:

$$\hat{p}_{\text{MAP}} = \hat{p}_{\text{EV}} = \frac{(\mu_a - \bar{y})}{K}, \text{ where } K = 1 + (\sigma_a^2/\, n\tau^2)$$

- The distribution of $\hat{p}$ is then

$$P(\hat{p}\,|\,p_{\text{parent}}) = \mathcal{N}(p_{\text{parent}}/K, \sigma_a^2/nK^2)$$

Where will $\hat{p}$ be relative to p_{parent}?

- Because $K > 1$ (for any values of σ_a, n, and τ), the expected value of p decreases with each generation:

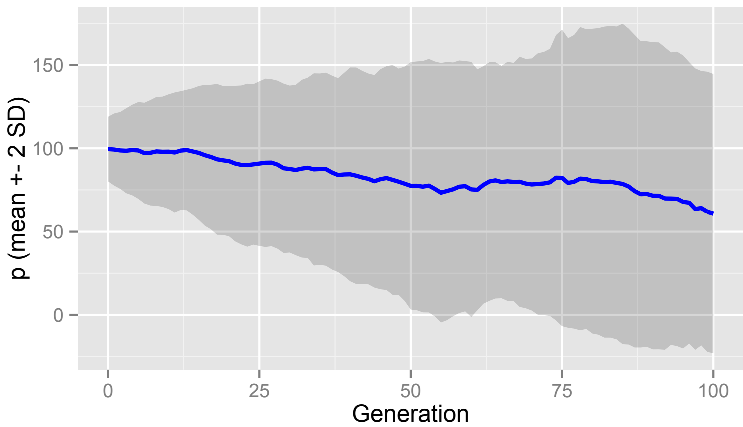
$$E(\hat{p}) = E(p_{\text{parent}})/K$$

- The variance of p moves over time towards the fixed point α_*:

$$\text{Var}(\hat{p}) = [\sigma_a^2/n + \text{Var}(p)]/K^2$$

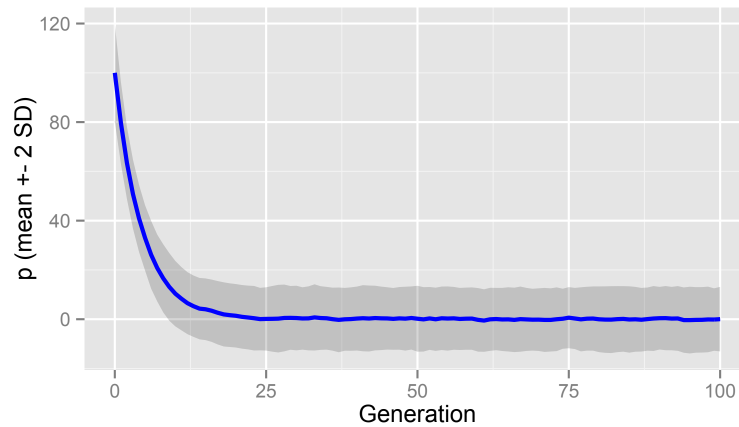
Evolution of π_t under $\mathcal{A}_{\text{simple}}, \mathcal{S}_{\text{single}}$

100 epochs, 100 examples, 1000 agents, $\tau = 100$



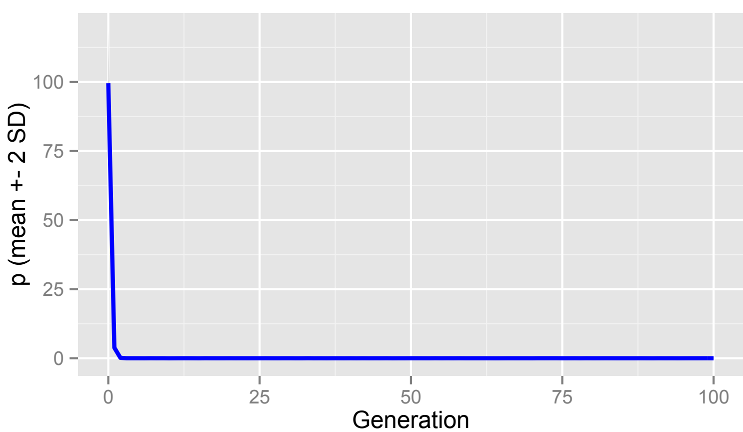
Evolution of π_t under $\mathcal{A}_{\text{simple}}, \mathcal{S}_{\text{single}}$

100 epochs, 100 examples, 1000 agents, $\tau = 10$



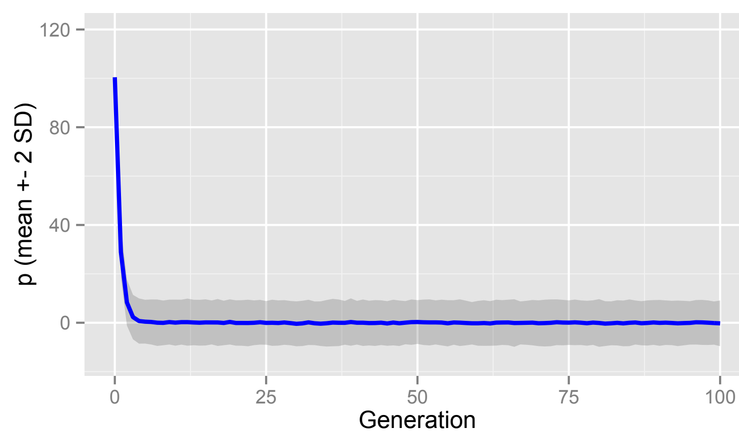
Evolution of π_t under $\mathcal{A}_{\text{simple}}, \mathcal{S}_{\text{single}}$

100 epochs, 100 examples, 1000 agents, $\tau = 1$



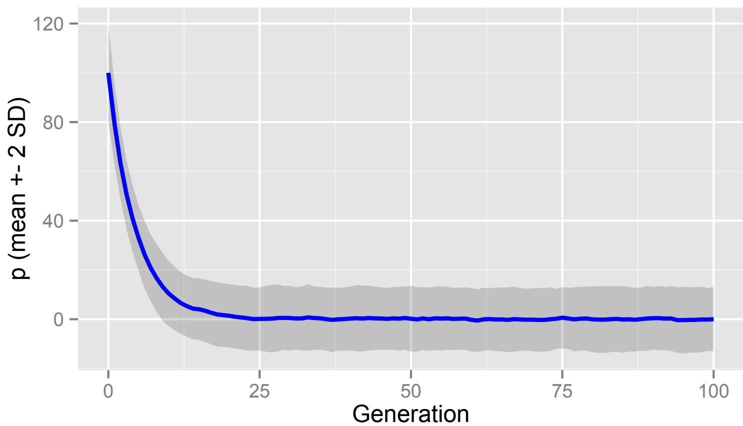
Evolution of π_t under $\mathcal{A}_{\text{simple}}, \mathcal{S}_{\text{single}}$

100 epochs, 10 examples, 1000 agents, $\tau = 10$



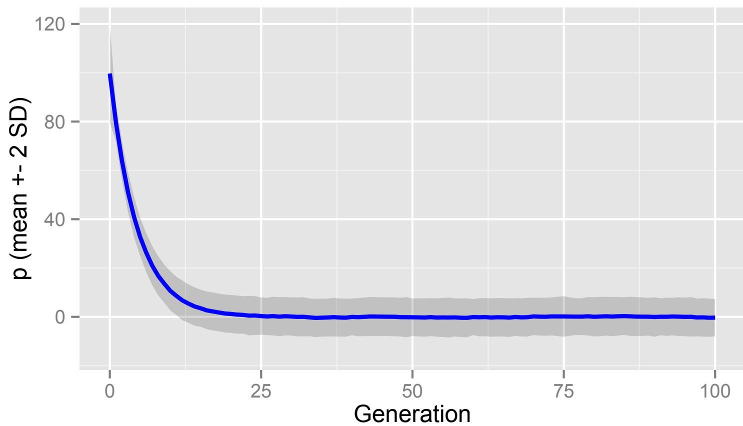
Evolution of π_t under $\mathcal{A}_{\text{simple}}, \mathcal{S}_{\text{single}}$

100 epochs, 100 examples, 1000 agents, $\tau = 10$



Evolution of π_t under $\mathcal{A}_{\text{simple}}, \mathcal{S}_{\text{multiple}}$

100 epochs, 100 examples, 1000 agents, $\tau = 10$

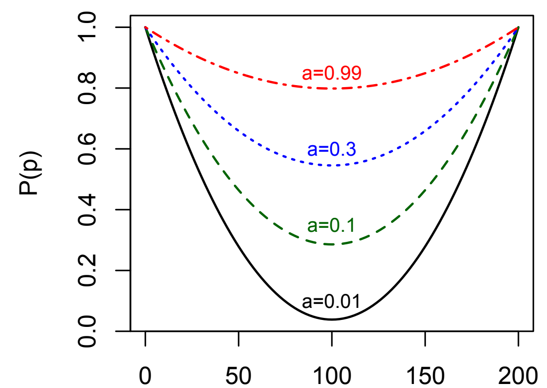


$\mathcal{A}_{\text{complex}}$

- Neither $\mathcal{A}_{\text{naive}}$ nor $\mathcal{A}_{\text{simple}}$ are adequate: one predicts constant change, the other lacks the ability to model *any* change
- Prior needs to encode some kind of category preference
- One option: quadratic polynomial with a minimum at $(\mu_a - \mu_b)/2$, concave up between 0 and $\mu_a - \mu_b$

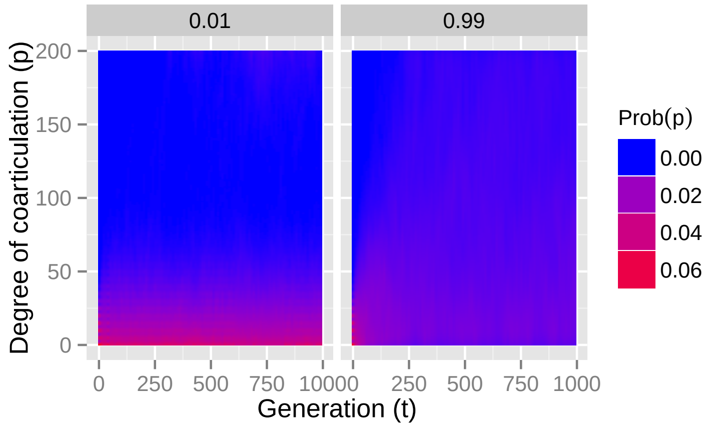
$\mathcal{A}_{\text{complex}}$

$$P(p) \propto \left[a(\mu_a - \mu_b)^2 + (p - (\mu_a - \mu_b)/2)^2\right]$$

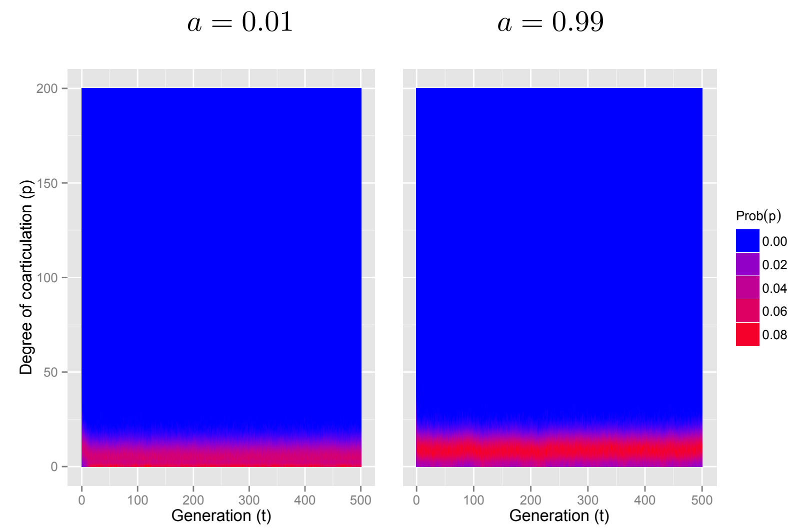


Evolution of π_t under $\mathcal{A}_{\text{complex}}, \mathcal{S}_{\text{single}}$

Evolution of density of p over time (indicated by color) with (left) a strong polynomial prior ($a = 0.01$) or (right) a weak polynomial prior ($a = 0.99$).

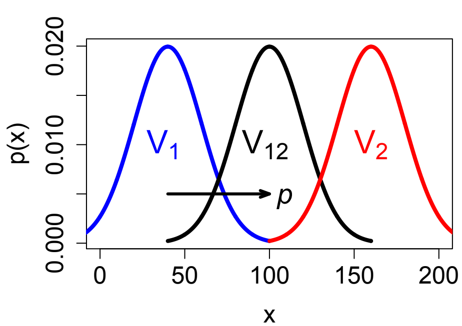


Evolution of π_t under $\mathcal{A}_{\text{complex}}, \mathcal{S}_{\text{multiple}}$



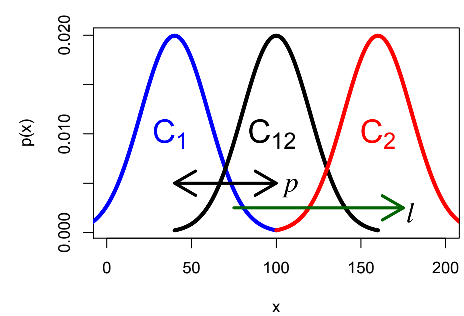
Evolution of π_t under $\mathcal{A}_{\text{complex}}$

- Stability is now possible, but change is still elusive – why?
- Prior is strong enough to bias learners towards p = 0 or $p = \mu_{V_1} - \mu_{V_2}$, but no bias towards full coarticulation.

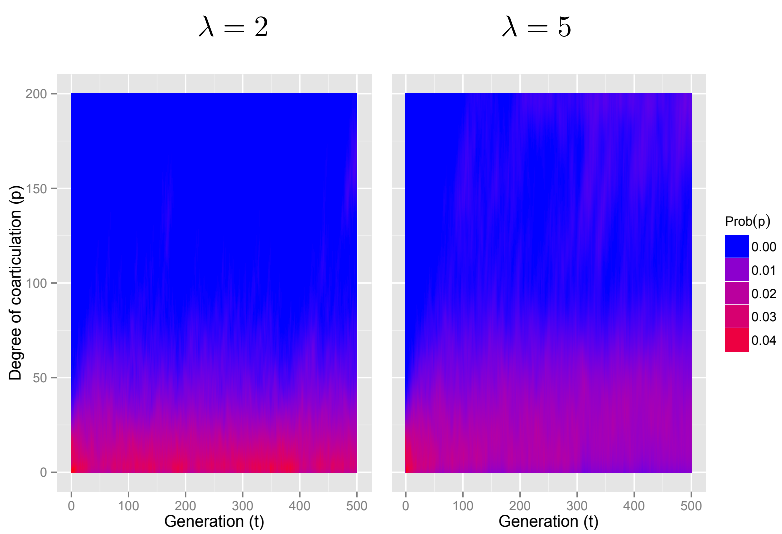


Evolution of π_t under $\mathcal{A}_{\text{complex}}$

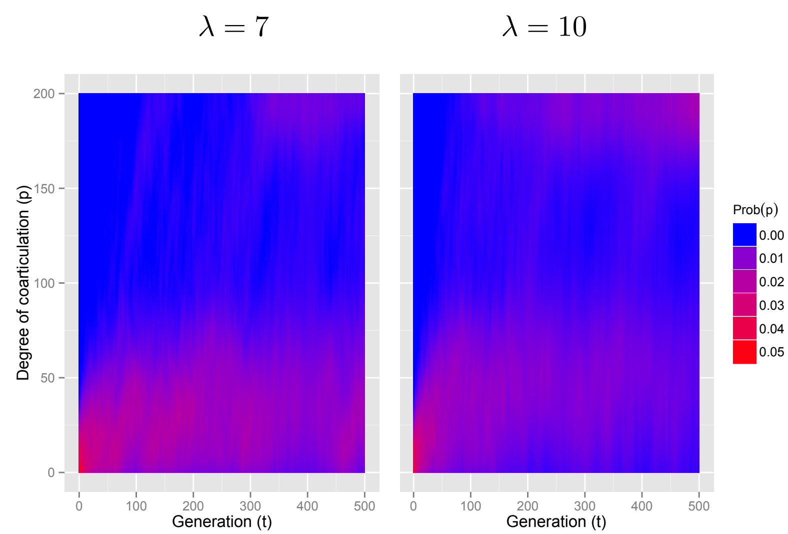
- One type of bias: external force increasing the likelihood of coarticulated variants (Pierrehumbert, 2002; Wedel, 2006)
- Here: assume some percentage (10%) of examples have been moved towards μ_{V_2} by a quantity $\ell \sim \mathcal{N}(\lambda, \lambda/2)$ (...)



Varying λ under $\mathcal{A}_{\text{complex}}, \mathcal{S}_{\text{single}}, a = 0.01$



Varying λ under $\mathcal{A}_{\text{complex}}, \mathcal{S}_{\text{single}}, a = 0.01$

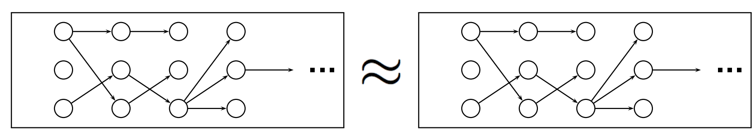


Discussion

- Bifurcation occurred (maybe?) as a system parameter (the amount of production bias) was varied past a critical value.
- Bifurcations suggested as a potential mechanism underlying actuation of linguistic change, but only shown to occur in models of change in discrete parameters (e.g. Niyogi 2006)
- Possibility of bifurcations in continuous case suggests they play a key role in the actuation of change more generally...?

Discussion

- Among the models considered here, empirically adequate account seems to require **both** a strong learning bias **and** some kind of trigger/production bias
- Assumptions about populations did not matter so much: impacted rates, but not qualitative outcomes



- (Note that this was not true for $\mathcal{A}_{\text{naive}}$.)

Possible extensions

- $\mathcal{A}_{\text{complex}}$: mapping the relation between a and ℓ, biasProp...
- $\mathcal{A}_{\text{simple}}$: lognormal, etc. prior; also prior + lenition
- Vary n_{teachers} from 1, 2, . . . many
- Less weird λ distribution
- Horizontal transmission
- Population subsets/network structure
- Distributions over multiple cues (hard)