A model of population dynamics applied to phonetic change

James Kirby (j.kirby@ed.ac.uk)
School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh EH8 9AD UK

Morgan Sonderegger (morgan.sonderegger@mcgill.ca)
Department of Linguistics, McGill University, Montreal, Quebec H3A 1A7 Canada

Abstract

We consider the problem of language evolution in a population setting, focusing on the case of continuous parameter learning. While theories of phonetic change tend to emphasize the types of transmission errors that could give rise to a shift in pronunciation norms, it is challenging to develop a model that allows for both stability as well as change. We model the acquisition of vowel-to-vowel coarticulation in both single- and multiple-teacher settings, considering progressively more restrictive prior learning biases. We demonstrate that both stability and change are possible at the population level, but only under fairly strong assumptions about the nature of learning and production biases.

Keywords: Language evolution; sound change; computational modeling; phonetics; coarticulation

Introduction

The problem of language evolution and change has received increased attention from a computational perspective in recent years (e.g. Niyogi & Berwick, 1995; Wedel, 2006; Kirby, Dowman, & Griffiths, 2007). Most of this work has focused on modeling either lexical or syntactic change, where the task is usually cast as deciding between competing *discrete* representation, e.g. different grammars (Baker, 2008). A similar approach is often taken in models of the evolution of sound patterns, where the learning problem is cast as one of deciding between discrete pronunciation variants (e.g. Niyogi, 2006).

However, learning a sound pattern of a language also consists of learning *continuous* phonetic cue distributions that describe how the sounds of that language are realized. Understanding the dynamics of these distributions is important for understanding sound change, because the seeds of category-level change are often claimed to be based in continuous phonetic variation (Ohala, 1993). In this paper, we address the evolution of sound patterns by considering the acquisition of continuous parameter distributions in a population setting. While we consider the particular example of a phonetic parameter, the basic results are applicable to the learning of continuous parameters more generally.

Stability and change in phonetic realization

In all languages, when a sound is produced in a connected stream of speech, its phonetic realization is influenced by the preceding and/or following context. This contextual variability, termed *coarticulation*, has often been argued to underlie a wide variety of sound changes in the world's languages. One example is a historical process known as *primary umlaut*, attested in Old High German beginning c. 750 AD, in which short low /a/ was fronted and raised to /e/ when a high front vowel or glide occurred in the following syllable, e.g. *[gasti:] > /gesti/ 'guests' (modern German *Gäste*). It has been proposed that the roots of umlaut may be traced to vowel-to-vowel coarticulation (Iverson & Salmons, 2003); however, vowel-to-vowel coarticulation did not invariably result in umlaut. For example, even while primary umlaut was spreading throughout West Germanic, it is clear that it did not affect Gothic (Campbell, 1998:75). The umlaut example illustrates a more general point: the mere *presence* of a potential trigger does not imply that phonetic change is inevitable. Thus, any empirically adequate model of how the sound pattern of a language evolves must account for instances of *stability* as well as change (Weinreich, Labov, & Herzog, 1968).

Learning bias in phonetic change

An important body of research on phonetic change has focused on establishing the preconditions for change to occur in a single speaker-hearer (Ohala, 1993). Similarly, computational models of phonetic change have mostly considered individuals, focusing on how *biases* in learning or in speech production/perception impact whether or not change occurs (Pierrehumbert, 2001). However, even if a change were to obtain at the level of a single speaker, its spread in the speech community is far from inevitable: social and cultural factors may conspire to inhibit or enhance a change in the population at large. In addition, the dynamics of linguistic populations are complex: how assumptions about individual speakers play out in population dynamics can be surprisingly nontrivial and dependent on assumptions about population structure (Niyogi, 2006). For both reasons, the general plausibility of accounts of contextually-driven phonetic change, and what role channel and learning biases play, cannot be properly assessed until their dynamics at the population level are better understood.

This paper explores the effects of different assumptions about bias and population structure on the evolution of phonetic categories in a population, as applied to a simplified version of primary umlaut in Germanic. In particular, we consider six models of learning how /a/ is pronounced before a high vowel. Our aim is a model which satifies three goals:

1. *Stability* of limited coarticulation in the population, as in pre-Old High German
2. *Stability* of full coarticulation in the population (e.g. umlaut), as in Old High German
3. *Change* from stable limited coarticulation to stable full coarticulation.

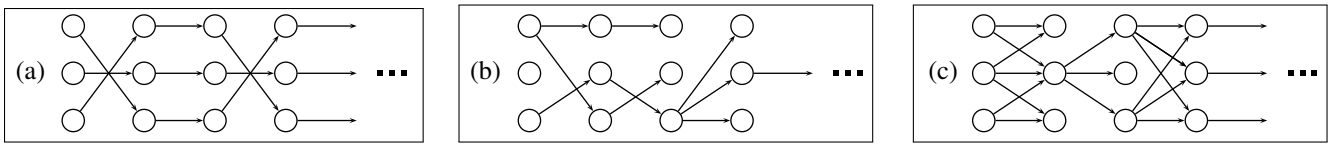

Figure 1: (a) Parallel diffusion chains (classic iterated learning). (b) Single-teacher scenario. (c) Multiple-teacher scenario.

Properties of populations

Fig. 1 illustrates three types of population structure. Fig. 1a shows a classic *iterated learning* (IL) scenario, also known as a *diffusion chain* (Smith, Kirby, & Brighton, 2003). In IL, each learner of generation $t+1$ receives input from a randomly chosen member of generation t. Thus, every member of a generation functions both as learner and as teacher. Fig. 1b illustrates *single-teacher learning with replacement*. This scenario differs crucially from classic IL in that while the input for a learner comes from exactly one teacher, some teachers may provide input to more than one learner, while others may not provide any. Finally, Fig. 1c illustrates *multiple-teacher learning with replacement*. Here, input may come from more than one teacher, although some teachers may not provide data to any leaners in the following generation.

The IL scenario has frequently been assumed in work on language evolution and change (e.g. Smith et al., 2003; Kirby et al., 2007).[1] While these models have well-understood dynamics and may be appropriate in some situations, in general different dynamics emerge in population learning scenarios (1b-c) (Dediu, 2009; Smith, 2009). In this work we focus on scenarios (1b-c), aiming to determine (1) what type of biases are necessary for stability and change to obtain in a population of learners, and (2) if and how such biases interact with differences in the number of teachers. We begin with a naive learning model (no prior) and then turn to consider progressively more restrictive priors.

Framework

We assume that (1) speech sounds have been organized into discrete segments; (2) the phonetic realisation of segments is subject to coarticulation; and (3) the learner has access to the complete segmental inventory. We consider here a simple language with the lexicon $\Sigma = \{V_1, V_2, V_{12}\}$, where V_{12} represents V_1 in the potentially coarticulation-inducing context of V_2. For the primary umlaut example, V_1 and V_2 can be thought of as the vowels /a/ and /i/ in isolation, and V_{12} as the vowel /a/ in a context where it is coarticulated towards /i/.

For simplicity, vowel tokens are represented by their first formant (F1) value, an acoustic measure of vowel height. We assume that the F1 distributions of V_1 and V_2 are known to all learners, are the same for all learners, and do not change over time. The distribution of V_{12} differs from that of V_1 only by an offset to the mean p, indicating how much V_1 is affected by coarticulation (i.e. raised) in the V_2 context. In all derivations below, we assume in particular that for a learner with parameter p, the three categories V_1, V_2, and V_{12} follow normal distributions in a single dimension (F1):

$$V_1 \sim N(\mu_a, \sigma_a^2),\ V_2 \sim N(\mu_i, \sigma_i^2),\ V_{12} \sim N(\mu_a - p, \sigma_a^2) \quad (1)$$

We assume that learners are divided into discrete generations $\mathcal{G}_t$ of size M. Each learner in generation $\mathcal{G}_t$ receives $n_{V_{12}}$ examples of V_{12} from the members of generation $\mathcal{G}_{t-1}$. The learner's task is then simply to infer p. The state of the population $\mathcal{G}_t$ can thus be characterized by the distribution $p \sim \pi_t(p)$. For simplicity, we assume that M is infinite, so the evolution of the population is not a stochastic process.

In $\mathcal{G}_{t+1}$, each learner is presented with n examples of V_{12} drawn from a sequence of teachers in $\mathcal{G}_t$ chosen by some sampling procedure $\mathcal{S}$.[2] Given these examples, the learner applies some learning algorithm $\mathcal{A}$. Assuming $\mathcal{S}$ and $\mathcal{A}$ are the same for all agents in $\mathcal{G}_{t+1}$, this implies the following evolution equation for π_t:

$$(\pi_{t+1}) = f_{\mathcal{S},\mathcal{A}}(\pi_t \,|\, \text{constants}) \quad (2)$$

For a given $\mathcal{A}$ and $\mathcal{S}$, our goal is to determine f, and characterize its behavior, in particular which (if any) of our modeling goals it satisfies.

Models

This section describes the evolutionary dynamics of a population of learners who estimate the degree of coarticulation from training data based on the assumption that these examples are independently and identically (i.i.d.) generated by a single source with a fixed p. We consider learners with three types of prior bias in estimating p, corresponding to three choices of $\mathcal{A}$: no prior ($\mathcal{A}_{\text{naive}}$), a simple prior ($\mathcal{A}_{\text{simple}}$), and a more complex prior ($\mathcal{A}_{\text{complex}}$). These learners are embedded in two of the types of populations shown in Fig. 1, corresponding to two choices of $\mathcal{S}$: (1b), in which a learner's input is provided by a single teacher ($\mathcal{S}_{\text{single}}$), and (1c), in which her input may be drawn from multiple teachers ($\mathcal{S}_{\text{multiple}}$).

$\mathcal{A}_{\text{naive}}$: Naive learning models

We first consider maximum-likelihood (ML) learners who are "naive" in the sense of having no prior over estimates of p.[3]

[1] Some work, e.g. Griffiths and Kalish (2007), has also considered models of type (1b), but only for the discrete parameter case.

[2] This is equivalent to sampling from $\pi_t(p)$ and generating an example from the distribution implied by the value of p chosen.

[3] These models are equivalent to special cases of 'blending inheritance' models of cultural evolution of a quantitative character (Boyd and Richerson, 1985: 71ff).

Model 1.1: Naive learning, single teacher First we consider a situation in which a learner in generation $\mathcal{G}_{t+1}$ receives examples from a single member of generation $\mathcal{G}_t$. Each learner is associated with a value p_{parent} (one draw from the π_t distribution, representing the single teacher's degree of coarticulation), which is used to generate n training examples $\vec{y} = (y_1, \ldots, y_n)$. Let $\bar{y}$ be the mean of this sample. Each example is normally distributed, following (1): $P(y_i) = N(\mu_a - p_{\text{parent}}, \sigma_a^2)$. The sample's mean is also normally distributed, with the same mean and reduced variance:

$$P((y_1 + \cdots + y_n)/n \,|\, p_{\text{parent}}) = N(\mu_a - p_{\text{parent}}, \sigma_a^2/n) \quad (3)$$

Given $\bar{y}$, the learner's maximum-likelihood estimate of p is $\hat{p} = \mu_a - \bar{y}$. Thus, using Eq. 3, the distribution over values of $\hat{p}$ the learner could acquire given p_{parent} is:

$$P(\hat{p} \,|\, p_{\text{parent}}) = N(p_{\text{parent}}, \sigma_a^2/n) \quad (4)$$

We are interested in the evolution of the distribution π_t: that is, the marginal distribution of $\hat{p}$ as a function of the distribution of p_{parent}. Abbreviating p_{parent} as p, this is:

$$\begin{aligned}\pi_{t+1}(\hat{p}) &= \int \underbrace{P((\hat{p} \,|\, p)}_{\text{Eq. 4}} \pi_t(p)\, dp \\ &= \int \pi_t(p) \cdot N_{\hat{p}}(p, \sigma_a^2/n)\, dp \quad (5)\end{aligned}$$

To get a sense of the evolution of π_t, we can compute how its mean and variance change over time. Let p be the random variable distributed according to π_t, and $\hat{p}$ the same for π_{t+1}. The expected value of $\hat{p}$ is then:

$$\begin{aligned}E[\hat{p}] = \int \pi_{t+1}(\hat{p})\, \hat{p}\, d\hat{p} &= \int \left[\int \pi_t(p) \cdot N_{\hat{p}}(p, \sigma_a^2/n)\, dp \right] \hat{p}\, d\hat{p} \\ &= \int \pi_t(p) \underbrace{\left[\int N_{\hat{p}}(p, \sigma_a^2/n)\, \hat{p}\, d\hat{p} \right]}_{=E[\hat{p} \,|\, p] = p} dp \\ &= \int \pi_t(p)\, p\, dp = E[p] \quad (6)\end{aligned}$$

By a similar derivation for $E[\hat{p}^2]$, the variance of $\hat{p}$ can be shown to be:

$$\begin{aligned}\text{Var}(\hat{p}) &= E[(\hat{p} - E[\hat{p}])^2] = E[\hat{p}^2] - E[\hat{p}]^2 \\ &= \sigma_a^2/n + \text{Var}(p) \quad (7)\end{aligned}$$

Thus, the distribution of p in the $n+1^{\text{th}}$ generation has the same mean as in the n^{th} generation, but larger variance; i.e., the distribution becomes more diffuse with each generation.[4]

[4]This contrasts with the common statement that 'blending inheritance' *reduces* variance of a quantitative trait over time (Boyd and Richerson, 1985: 75). However, stable or increasing variance are possible for particular cases of Boyd and Richerson's model, such as the case considered here where each learner has a single 'cultural parent' and there is noise in estimating the parent's cultural model.

Model 1.2: Naive learning, multiple teachers We now consider the case where a learner in generation $\mathcal{G}_{t+1}$ receives each training example from a randomly-chosen teacher in generation $\mathcal{G}_{t+1}$. This is equivalent to drawing n values of p from π_t, $\vec{p} = (p_1, \ldots, p_n)$, and for each p_i generating one training example y_i:

$$P(y_i \,|\, p_i) = N(\mu_a - p_i, \sigma_a) \qquad i = 1, \ldots, n \quad (8)$$

As in the single-teacher case, we assume that the learner chooses the ML estimate for p, $\hat{p} = \mu_a - \bar{y}$. Using (8) and the fact that the y_i are independent and normally distributed:

$$\begin{aligned}&P(\bar{y} \,|\, \vec{p}) = N(\mu_a - (p_1 + \cdots + p_n)/n, \sigma_a^2/n) \\ \Longrightarrow\ &P(\hat{p} \,|\, \vec{p}) = N(\bar{p}, \sigma_a^2/n) \quad (9)\end{aligned}$$

where $\bar{p} = (p_1 + \cdots + p_n)/n$. Thus, the learner's estimate $\hat{p}$ is the mean of the p values which generated the training data, plus some noise.

To obtain $\pi_{t+1}(\hat{p})$, the marginal distribution of $\hat{p}$, we integrate out $p_1, \ldots, p_n$ from (9):

$$\pi_{t+1}(\hat{p}) = \int N_{\hat{p}}(\bar{p}, \sigma_a^2/n) \prod_{i=1}^{n} \pi_t(p_i)\, dp_i \quad (10)$$

As in the single-teacher case, we can get a sense of how the distribution of p evolves by computing the mean and variance of π_{t+1}. Let p_t be the random variable with distribution π_t. The expected value and variance of $\hat{p}$ can be shown to be:

$$E(\hat{p}) = E(p_t), \quad \text{Var}(\hat{p}) = \sigma_a^2/n + \text{Var}(p_t)/n \quad (11)$$

Some algebra shows that

$$(\text{Var}(\hat{p}) - \alpha_*) = (\text{Var}(p) - \alpha_*)/n$$

where $\alpha_* = \sigma_a^2/(n-1)$. The variance of the distribution of p moves over time towards α_*; if already at α_*, it stays there forever. Thus, the mean of the distribution of p stays the same over time, but its variance moves towards a single value.

Summary Whether the single-teacher (1.1) or multiple-teacher (1.2) scenario is assumed, the naive learning models predict that the average degree of coarticulation in the population will not change over time. It follows that, under these assumptions, change from little coarticulation to full coarticulation (Goal 3) is not possible.

The single-teacher model has an additional problem. The variability of the degree of coarticulation in the population is predicted to increase with each generation., i.e. speakers come to coarticulate increasingly differently. Intuitively, because each production is noisy, the learner's estimate of the degree of coarticulation is *inherently noisy* (Eq. 4): it is impossible to exactly acquire the target value of the parent from a finite sample. Increasing population-level variation in the degree of coarticulation over time is clearly empirically inadequate, because the effects of umlaut are generally either present or absent in a given population. Thus, Model 1.1 does not allow for stability of a population with little coarticulation (Goal 1) or full coarticulation (Goal 2).

$\mathcal{A}_{\text{simple}}$: Simple prior models

Intuitively, the reason that the single-teacher naive learning model fails to allow for stability around a particular value of p is that there is no force counteracting the noise in each learner's estimate (4), which causes the distribution of p values to spread out over time. In this section, we consider the effect of a simple prior learning bias on the evolution of p.

As above, we assume the learner estimates p based on the assumption that data is generated i.i.d. from a source with a fixed p. The distribution of the data under this assumption is

$$P(\vec{y}\,|\,p) = P(y_1\,|\,p)\cdots P(y_n\,|\,p) \tag{12}$$

$$= \frac{\exp\left[-\sum_{i=1}^{n}(y_i-(\mu_a+p))^2/(2\sigma_a^2)\right]}{(2\pi\sigma_a^2)^{n/2}} \tag{13}$$

However, we now assume that learners' knowledge about p is probabilistic: they begin with a prior distribution ($P(p) = \alpha(p)$) on how likely different values of p are *a priori*, which is updated to a posterior distribution based on the data ($P(p\,|\,\vec{y})$).

Recall that the population of naive learners from a single teacher did not show the simplest possible empirically adequate dynamics: stability of the distribution of p over time near $p = 0$; i.e., most people have a minimal (but fixed) degree of coarticulation. As a first pass to see if this behavior is possible with learners who reason probabilistically about p, we assume learners have a prior biasing them towards values of p near 0, with values away from 0 becoming increasingly less likely. Intuitively, this prior "should" sharpen the distribution of p towards $p = 0$ over time, counteracting the effect of transmission noise which tends to make the distribution of p spread out more in each generation (as in Model 1.1).

For simplicity, we assume a Gaussian prior $\alpha \sim N(0,\tau^2)$. The posterior is then simply:

$$P(p\,|\,\vec{y}) = P(\vec{y}\,|\,p)P(p)/P(\vec{y}) \tag{14}$$

The learner must pick a point estimate of p, denoted $\hat{p}$, using $P(p\,|\,\vec{y})$. The two familiar strategies, choosing the maximum or expected value of the posterior (abbreviated MAP, EV), turn out to be equivalent:

$$\hat{p}_{\text{MAP}} = \hat{p}_{\text{EV}} = \frac{(\mu_a-\bar{y})}{1+\sigma_a^2/\,n\tau^2} \tag{15}$$

Abbreviating the denominator of (15) as $K = 1+\sigma_a^2/(n\tau^2)$, these estimates of p may then be equivalently written as $\hat{p} = (\mu_a-\bar{y})/K$.

As above, we can now consider the consequences of this learning strategy under different population scenarios.[5]

Model 2.1: Simple prior, single teacher We again first assume a scenario in which each learner in generation $\mathcal{G}_{t+1}$ receives n training examples from a single member of generation $\mathcal{G}_t$, who has coarticulatory parameter p_{parent}, abbreviated as p. The distribution of p is $P(p) = \pi_t(p)$.

[5]The ML estimate of $\hat{p}$ in the no-prior case above, $\mu_a-\bar{y}$, can thus be thought of as the Gaussian-prior estimate when the prior is very flat relative to the dispersion of the phonetic category ($\tau \gg \sigma_a$).

We first determine the distribution of a learner's estimate $\hat{p}$, given fixed p and data $\vec{y}$. $\bar{y}$ is normally distributed, as described by (3), as in the no-prior case. Because $\bar{y}$ is normally distributed and $\hat{p} = (\mu_a-\bar{y})/K$, the distribution of $\hat{p}$ is

$$P(\hat{p}\,|\,p_{\text{parent}}) = N(p_{\text{parent}}/K, \sigma_a^2/nK^2) \tag{16}$$

Thus, on average, the learner's estimate of p is closer to 0 than the parent's value.

We can now compute the marginal distribution of $\hat{p}$:

$$\pi_{t+1}(\hat{p}) = P(\hat{p}) = \int \underbrace{P(\hat{p}\,|\,p)}_{(16)} P(p)\,dp \tag{17}$$

$$= \int N_{\hat{p}}(p/K, \sigma_a^2/nK^2)\pi_t(p)\,dp \tag{18}$$

As in the no-prior case, we can gain some understanding of the evolution equation (18) by examining how the expectation and variance of p evolve. By a similar derivation to (6), it can be shown that the expectation of $\hat{p}$ is:

$$E(\hat{p}) = E(p_{\text{parent}})/K \tag{19}$$

Because $K > 1$ (for any values of σ_a, n, and τ), the expected value of the coarticulation parameter *decreases* with each generation. By a similar derivation to the no-prior case, the variance of $\hat{p}$ can be shown to be

$$\text{Var}(\hat{p}) = [\sigma_a^2/n + \text{Var}(p)]/K^2 \tag{20}$$

and some algebra shows that

$$(\text{Var}(\hat{p}) - \alpha_*) = (\text{Var}(p) - \alpha_*)/K^2$$

where $\alpha_* = \frac{\sigma_a^2}{n(K^2-1)}$. Because $K > 1$, the variance of p moves over time towards the fixed point α_*, as in Model 1.2. Thus (as noted by Smith, 2009 in other settings), the distribution of coarticulation in the population does not converge to the prior, unlike the well-known result of Griffiths and Kalish (2007).

Model 2.2: Simple prior, multiple teachers The situation is similar under $\mathcal{S}_{\text{multiple}}$. The mean of p can be shown to evolve exactly as in the single-teacher case (19), towards 0. Similarly, the variance looks very similar to the evolution in the single-teacher case (20), except for an extra factor of n in the denominator. The variance again evolves towards a fixed point, now $\alpha_* = \sigma_a^2/(nK^2-1)$, but in this case more quickly than in Model 2.1. Intuitively, this means that because learners have a strong prior against coarticulation, evidence for coarticulation at the level of the individual is mitigated and is unlikely to spread throughout the population.

Summary For both single- and multiple-teacher scenarios, a simple Gaussian prior drives the value of p to 0, predicting phonologization of coarticulation to be impossible. Thus, both Model 2.1 and 2.2 meet modeling Goal 1 (stability of little coarticulation), but neither of Goals 2 or 3.

$\mathcal{A}_{\text{complex}}$: Complex prior models

The preceding section has shown that the distribution of p in populations of learners with a Gaussian prior always converges to 0. This simple prior model is empirically inadequate, because it fails to predict the possibility of stable coarticulation in a population. We therefore considered several more complex priors. Here we discuss one such prior, a quadratic polynomial with a minimum at $(\mu_a - \mu_i)/2$ which is concave up between 0 and $\mu_a - \mu_i$:

$$P(p) \propto \left[a(\mu_a - \mu_i)^2 + (p - (\mu_a - \mu_i)/2)^2\right] \qquad (21)$$

Here, a is a scale parameter controlling the "strength" of the prior: as $a \to 0$, values of p near the endpoints are maximally weighted relative to values near $(\mu_a - \mu_i)/2$; as a is increased, the prior is progressively flatter.

We assume the learner takes the MAP estimate of p for values of p in $[0,\ \mu_a - \mu_i]$, which does not have a closed form solution, but can be found numerically. We thus proceeded by simulation to determine the evolution of the distribution of p over time in this case. The results reported here assume $\mu_a - \mu_i = 200$ and a strong prior ($a = 0.01$) in a multiple-teacher setting. We used large generation sizes ($M = 10000$) to approximate the deterministic behavior of infinite populations. Due to space constraints we discuss only the results for multiple-teacher models; the results for analogous single-teacher models are similar, in terms of our modeling goals.

Model 3.1: Complex prior First, we considered cases where there is little coarticulation in the population ($p_0 \sim N(10, 10)$) and where primary umlaut is effectively complete ($p_0 \sim N(190, 10)$). The evolution of density estimates for p over 1000 generations can be seen in the first panel of Fig. 2. While there is a slight shift in the mean and variance, they reach relatively stable values by around 1000 generations. However, the strength of the prior, in terms of the value of a, is important: as seen in the second panel of Fig. 2, for a weak prior ($a = 0.99$), the variance of p in the population increases quickly over time, with results similar to the no-prior case discussed above.

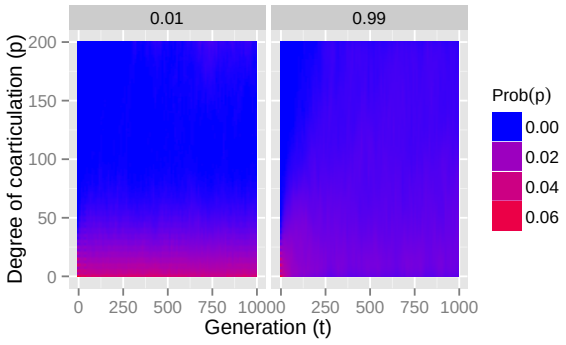

Figure 2: Evolution of density of p over time (indicated by color) with (left) a strong polynomial prior ($a = 0.01$) or (right) a weak polynomial prior ($a = 0.99$).

The simulation results suggest that a strong polynomial prior can result in a stable distribution for p in the population over time, with most learners having values near 0 (little coarticulation) or 200 (full coarticulation). Model 3.1 thus satisfies our Goals 1 and 2.

Model 3.2: Complex prior, production bias More extensive simulation with Model 3.1, however, suggests that it does not satisfy Goal 3: it is never possible for a population to transition from a stable state of little articulation to a stable state of full coarticulation. The reason is intuitively clear: the prior is strong enough to bias learners towards *either* $p = 0$ or $p = \mu_a - \mu_i$, but there is no countervailing force which could bias learners towards full coarticulation.

One plausible type of bias is an external force that increases the likelihood of coarticulated variants. Here, we implement a systematic production bias by assuming that some percentage of the learner's data have been moved towards μ_i by a quantity $\ell \sim N(\lambda, \lambda/2)$; that is, they are coarticulated *more* than expected from the teacher's value for the coarticulatory parameter. This kind of bias, corresponding to a general tendency in speech production to over- or undershoot articulatory targets, is commonly considered in computational models of phonetic change (e.g. Pierrehumbert, 2001).

Assuming a strong polynomial prior $a = 0.01$, change from no to full coarticulation turns out to indeed be possible, but only for a sufficiently large bias. Fig. 3 illustrates this with bias factors $\lambda = 2$ and $\lambda = 10$, starting in a state with little coarticulation ($p_0 = 10$), in which 10% of tokens in each generation were subject to a lenition bias. As in Model 3.1, for a strong enough prior (low a) with *no* bias, the little-coarticulation state is stable. As the amount of bias (λ) is increased past a critical value, there is a rapid shift of the population to a stable state where most learners have full coarticulation. That is, there is a *bifurcation* where the amount of bias has overcome the stabilizing affect of the prior, and the little-coarticulation state becomes unstable. These results also illustrate a more general tradeoff between the strength of the prior and the amount of bias observed in further simulations (not shown here): for a stronger prior, the critical value of λ increases: more bias is needed to overcome the prior.

Thus, Model 3.2 meets all three of our modeling goals: (1) a stable population with little coarticulation, (2) a stable population with full coarticulation, and (3) a rapid transition from little to full coarticulation are all possible, for particular initial conditions and values of the system parameters (a, λ).

Discussion

Our main goal in this paper was to evaluate how assumptions about bias and population structure for a population of learners of a continuous parameter translated into population-level models capable of modeling three empirically-observed scenarios of stability and change.

One interesting result was that population structure did not necessarily have much effect on the dynamics. For the naive learning scenario, the single-teacher and multiple-

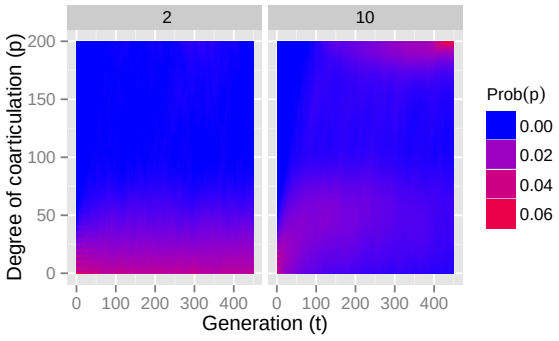

Figure 3: Evolution of density of p over time (indicated by color) with a strong polynomial prior ($a = 0.01$), with 10% of tokens subject to bias factor $\lambda = 2$ (left) or $\lambda = 10$ (right).

teacher models had qualitatively different dynamics: the variance in the degree of coarticulation stabilized over time in the multiple-teacher model, but *increased* over time in the single-teacher model. But for the simple prior and complex prior models, whether a single-teacher or multiple-teacher scenario was assumed largely impacted the rate at which a stable state was reached, rather than changing the qualitative outcome. Given that social structure plays a key role in the actuation and spread of language change (Labov, 2001), future work should further explore the role of different population structures with more complex teacher-learner relations.[6]

On the other hand, assumptions about bias mattered a great deal. When no or weak learning bias ($\mathcal{A}_{\text{naive}}$, or $\mathcal{A}_{\text{simple}}$ with high a) was assumed, stability of the distribution of the coarticulatory parameter p in the population was impossible. When a strong bias towards non-coarticulation was assumed ($\mathcal{A}_{\text{naive}}$ with low a), stability of minimal coarticulation was possible (Goal 1), but stability of full coarticulation and change between the two (Goals 2, 3) were not. It was only after assuming learners have a strong prior biasing them towards *either* little or full coarticulation, along with introducing an explicit unidirectional pressure to coarticulate, that it was possible to have primary umlaut: change (Goal 3) from stability of little coarticulation (Goal 1) to stability of full coarticulation (Goal 2).

Model 3.2, which met all three goals, shows a *bifurcation*: change from one stable state (little coarticulation in the population) to another (full coarticulation in the population) occurred suddenly as a system parameter (the amount of production bias) was varied past a critical value. Bifurcations in linguistic populations have been suggested as a potential mechanism underlying the actuation of linguistic change, but to our knowledge have previously only been shown to occur in models of change in *discrete* parameters (e.g. Niyogi, 2006). Our demonstration that bifurcations are possible in a population of learners of a *continuous* parameter supports the hypothesis that bifurcations play a key role in the actuation of language change more generally. Future work should explore whether such bifurcations emerge in models that more accurately reflect the social structure of speech communities, and where the outcome of learning is a distribution over multiple phonetic cues, rather than a single cue.

[6] A additional extension to be explored is horizontal transmission. In the present models, learners do not receive input from members of their own generation, but this could impact the dynamics as well.

References

Baker, A. (2008). Computational approaches to the study of language change. *Language and Linguistics Compass*, *2*(2), 289–307.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Campbell, L. (1998). *Historical linguistics: An introduction*. Cambridge, MA: MIT Press.

Dediu, D. (2009). Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise? *Journal of Theoretical Biology*, *259*, 552–561.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441–480.

Iverson, G. K., & Salmons, J. C. (2003). The ingenerate motivation of sound change. In R. Hickey (Ed.), *Motives for language change.* Cambridge: Cambridge University Press.

Kirby, S., Dowman, M., & Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*, 5241–5245.

Labov, W. (2001). *Principles of linguistic change vol. 2: Social factors*. Oxford: Oxford University Press.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Niyogi, P., & Berwick, R. (1995). *The logical problem of language change* (AI Memo No. 1516). MIT.

Ohala, J. J. (1993). The phonetics of sound change. In C. Jones (Ed.), *Historical Linguistics: Problems and Perspectives.* London: Longman.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 697–702). Austin: Cognitive Science Society.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*, 371–386.

Wedel, A. (2006). Exemplar models, evolution and language change. *The Linguistic Review*, *23*, 247–274.

Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical foundations for a theory of language change. In W. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics.* Austin: University of Texas.