UNIVERSALITY IN CULTURAL TRANSMISSION

Bill Thompson

AI Lab, Vrije Universiteit Brussel
bill@ai.vub.ac.be

ABSTRACT

Many physical systems exhibit *universality*: system-level behaviour is invariant to differences in the micro-level interacting elements that make up the system. Here I explore the possibility that, broadly understood, this property may also be true of *cultural transmission*, the process through which languages gain and lose structure. I take a powerful but computationally expensive Bayesian model for unsupervised induction of phonetic categories – the *infinite mixture-of-Gaussians* model [3] – and adapt it to include a lightweight, psychologically plausible scheme for rapid approximate posterior inference. I use this model to simulate cultural transmission among populations of learners, and search for signatures of the insensitivity that underpins *universality*: I ask how the population-level distribution of phonetic categories varies – or not – as a function of the biases inherent in the model of category acquisition. I discuss the potential significance of this connection to physical systems, and argue that this principle of insensitivity may have interesting consequences for the evolution of phonetic systems.

Keywords: Cultural Transmission; Universality; Bayesian Inference; Phonetic Categories

1. INTRODUCTION

An astonishing number of physical systems can be said to exhibit the property of *universality*: macro-level behaviour of the system is invariant to variations in the details at the micro-level. For example, perhaps the most general case of universality lies in the fact that all physical systems in equilibrium are subject to the same laws of thermo-dynamics, regardless of the particular atomic interactions that underpin the system. Universality has been discovered in a remarkable range of contexts, from the critical exponents that describe phase-transitions in a range of molecularly diverse substances, to the balance of chaos and order in seemingly unconnected dynamical systems, ranging from hypothetical mathematical constructs like the Zeta function [7], to the unregulated bus networks of Cuernavaca, Mexico [5]. Universality has also been argued to be a principle of broad significance to multi-agent dynamical systems in AI [9], when understood to describe: "...any system of interacting elements whose qualitative or quantitative system-level behaviour includes characteristics that are invariant under changes in the individual behaviour and detailed interaction of the elements." [9, pp.2].

Here I adopt this perspective and explore the possibility that, in some important respects, cultural transmission may also exhibit dynamics reminiscent of universality. In particular, I suggest that under some conditions cultural transmission may exhibit universality with respect to characteristics of the inductive biases that underpin learning. In section 2, I describe a Bayesian model for unsupervised inference of phonetic categories that permits a flexible range of inductive biases. In section 3 I simulate cultural transmission under this model, and show that there are conditions where differences in these biases do not result in differences at the population-level distribution of phonetic categories. Finally, section 4 lays out some potential consequences for our understanding of the evolution of phonetic capacities.

2. A MODEL OF PHONETIC CATEGORY ACQUISITION

Here I lay out a Bayesian model of phonetic category learning grounded in statistical inference over distributional cues. Following e.g. [2], I adopt the *mixture of Gaussians* representation of phonetic categories: in particular, I adapt the *infinite mixture of Gaussians* (iMOG) model developed by [3], and reformulate important aspects to reflect a simple, lightweight algorithm for posterior inference.

Models of phonetic category learning based on distributional statistics have gained considerable attention recently [6], and have proven particularly useful tools for exploring difficult acquisition problems, such as the rich statistical dependencies that could allow lexical statistics to bootstrap phonetic category acquisition [3], or the latent hierarchical structure that allows generalisation between phonetic categories [8], for example. This trend is in line with the broader movement to explore domain-

independent rational statistical inference as an explanation for inductive leaps that have traditionally been thought to indicate specialised inductive biases, particularly with respect to language e.g. [10]. Though the promise of these models is clear, there remain many fascinating open questions concerning the psychological mechanisms responsible for approximating the prohibitively complex computations that underpin these inferential models.

Here I chip away at the considerable computational resources required to implement unsupervised phonetic category acquisition in the iMOG while maintaining its desirable qualitative properties. In line with comparable variations on this model for of category acquisition *en général* [11], I simplify the model by implementing a psychologically plausible, sequential, greedy algorithm for approximating the posterior distribution over category assignments it implies, and by assuming a MAP point-estimate approach to parameter estimation for individual categories.

2.1. Phonetic Categories as Gaussian Distributions

Here I adopt the abstraction that a single acoustic feature f is relevant to the classification of speech sounds into categories. For example, f could be understood as voice onset time, or an absolute formant value, used in a language to distinguish one phonetic category from another. The learner observes a sequence of unlabelled (the learner does not know which category each sound represents) phonetic tokens $X = (x_1, \ldots, x_N)$, where each observation x_i is a speech sound exhibiting a value for f from a continuous range, for $i = 1 \ldots, N$. The model assumes that the learner's goal is to assign each observation x_i to a phonetic category c.

The major value of the iMOG is that it allows the number, C, of underlying phonetic categories to be inferred directly from the data, without being specified in advance or subject to an upper limit. Let $Z = (z_1, \ldots, z_n)$ be a partition on X, such that $z_i = c$ is an index which specifies that observation x_i has been assigned to category c. Each phonetic category c is assumed to be characterised by a Gaussian distribution with mean μ_c and variance σ_c^2. This allows a simple form for the likelihood that sound x_i would be produced from category c:

$$\Pr(x_i|\mu_c, \sigma_c^2) = \mathcal{N}(x_i; \mu_c, \sigma_c^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \quad (1)$$

The learner must infer the Gaussian parameters describing the distribution of sounds that will be produced from a category: she must induce estimates $\hat{\mu}_c$ and $\hat{\sigma}_c^2$. Again assuming Bayesian inference for these parameters, we must specify a prior $\Pr(\mu_c, \sigma_c^2)$ that captures the learner's initial expectations about μ_c and σ_c^2.

2.2. Prior Over Category Means and Variances

A natural conjugate prior for these parameters is the normal-inverse-chi-squared distribution. Dropping the index c for notational convenience, the prior is defined to be:

$$\begin{aligned}\Pr(\mu, \sigma^2|\Lambda) &= \mathcal{N}\mathcal{I}\chi^2(\mu, \sigma^2; \mu_0, k_0, \nu_0, \sigma_0^2) \\ &= \mathcal{N}(\mu; \mu_0, \sigma^2/k_0)\chi^{-2}(\sigma^2; \nu_0, \sigma_0^2) \\ &= \frac{1}{\beta}\sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)}\exp(\gamma) \qquad (2)\end{aligned}$$

$$\beta = \frac{\sqrt{2\pi}}{\sqrt{k_0}}\Gamma(\nu_0/2)\left(\frac{2}{\nu_0\sigma_0^2}\right)^{\nu_0/2} \qquad (3)$$

$$\gamma = -\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + k_0[\mu_0 - \mu]^2\right), \qquad (4)$$

where $\Lambda = (\mu_0, k_0, \sigma_0^2, \nu_0)$ are the parameters of the prior. Though the expression for the prior is necessarily complex (it must dictate a probability density function over two independent continuous variables), this formulation affords some useful mathematical properties, and has an intuitive interpretation: μ_0 is the initial guess for μ_c, and k_0 defines the learner's confidence in that guess; σ_0^2 is the initial guess about the variance σ_c^2, and likewise ν_0 determines the confidence in that guess.

2.3. Estimating μ_c and σ_c^2

Given a set of sounds $X_c = \{x_i : z_i = c\}$ believed to have been generated from category c, the learner combines this data with her prior beliefs to arrive at updated beliefs which are captured by the *posterior* distribution:

$$\Pr(\mu_c, \sigma_c^2|X_c, \Lambda) \sim \Pr(X_c|\mu_c, \sigma_c^2)\Pr(\mu_c, \sigma_c^2|\Lambda), \quad (5)$$

where the data likelihood factors into the product of the individual speech sounds:

$$\Pr(X_c|\mu_c, \sigma_c^2) = \prod_{x_i \in X_c} \Pr(x_i|\mu_c, \sigma_c^2). \qquad (6)$$

In contrast to [3], I assume that, for a given category, the learner induces estimates $\hat{\mu}_c$ and $\hat{\sigma}_c^2$ of the mean and variance of the Gaussian distribution for that category that reflect the *maximum a posteriori* (MAP) point-estimates, which maximise

$\Pr(\mu_c, \sigma_c^2 | X_c, \Lambda)$: these are the Bayesian equivalent of maximum-likelihood estimates, and have known closed-form expressions that don't require heavy integration over the cumbersome 2-dimensional posterior:

$$\hat{\mu_c} = \frac{\mu_0 k_0 + N_c \bar{X}_c}{k_n} \tag{7}$$

$$\hat{\sigma}_c^2 = \frac{v_n \sigma_n^2}{v_n - 1}\,, \tag{8}$$

where $\bar{X}_c$ represents the sample mean of X_c; $N_c = |X_c|$ is the number of speech sounds assosciated with category c; $v_n = v_0 + N_c$; $k_n = k_0 + N_c$; and:

$$\sigma_n^2 = \frac{1}{v_n}\left(v_0 \sigma_0^2 + (N_c - 1) \sum_{x_i \in X_c} (x_i - \bar{X}_c)^2 + \frac{N_c k_0}{N_c + k_0} [\mu_0 - \bar{X}_c]^2 \right). \tag{9}$$

2.4. Prior Over Number of Categories

The iMOG assumes a Dirichlet process (DP) prior over possible partitions of the data into category assignments Z. The DP has two parameters: α, a *concentration parameter* which implements a bias to hypothesise more ($\alpha \to \infty$) or fewer ($\alpha \to 0$) underlying categories to explain the data; and G_0, a *base distribution* which in this case provides the prior over the parameters μ_c and σ_c^2 for individual categories $c = 1, \ldots, \infty$. Here, the normal-inverse-chi-squared distribution described in equation (2) acts as the base distribution over the kinds of categories learners are likely to encounter, and α is a parameter I will vary.

Full details of the complete statistical model can be found in [3]. However, here I leverage the fact that the DP can formulated as sequential process specifying a form for the categorisation decisions to be made upon encountering a new data-point. When deciding upon a category assignment for speech sound x_i, the learner is computing:

$$\Pr(z_i = c | x_i, Z_{\_i}) = \Pr(x_i | z_i = c, Z_{\_i}) \Pr(z_i = c | Z_{\_i})\,, \tag{10}$$

where $Z_{\_i}$ represents the existing category assignments, and $\Pr(x_i | z_i = c, Z_{\_i}) = \Pr(x_i | \hat{\mu}_c, \hat{\sigma}_c^2)$ is the likelihood of observing speech sound x_i under this category given the sounds currently assigned to it, and the associated current estimate of its parameters, following eq. (1), with $\hat{\mu}_c$ and $\hat{\sigma}_c^2$ substituted for μ_c and σ_c^2 respectively. The sequential formulation of the DP, sometimes referred to as the *Chinese restaurant process*, allows us to formulate the prior over partitions into category assignments as a probabilistic choice between assigning a new observation to an existing category ($N_c \geq 1$) or creating a new category ($N_c = 0$):

$$\Pr(z_i = c | Z_{\_i}) = \begin{cases} N_c \Big/ \Big[\sum_{c=1}^{C} (N_c) - 1 + \alpha \Big], \text{ if } N_c \geq 1 \\ \alpha \Big/ \Big[\sum_{c=1}^{C} (N_c) - 1 + \alpha \Big], \text{ if } N_c = 0\,. \end{cases} \tag{11}$$

2.5. Rapid Sequential Approximate Posterior Inference

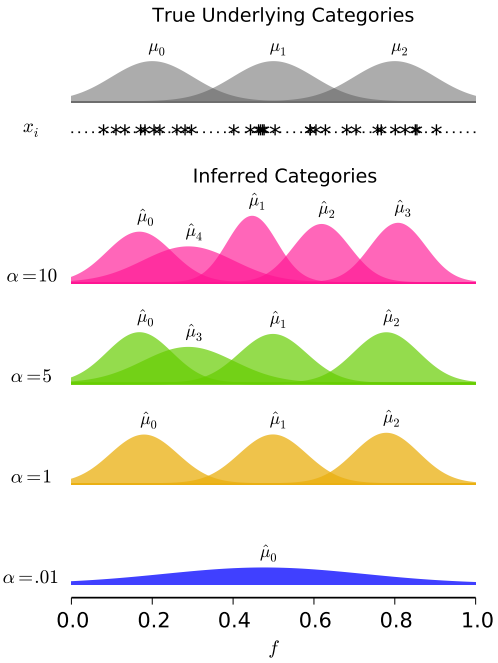
Finally, and perhaps most crucially, I assume a sequential, greedy updating scheme that provides a simple, psychologically plausible approximation to the posterior distribution over possible partitions Z. Specifically, I assume the learner arrives at a single fixed set of category assignments $\mathbf{Z}^*$ by sequentially choosing, upon receipt of each new observation x_i, to assign that data-point to whichever category maximises the posterior probability of assignment (eq. (10)) given the existing set of assignments, so that:

$$\mathbf{z}_\mathbf{i}^* = \arg\max_c \Pr(z_i = c | x_i, Z_{\_i}) \tag{12}$$

The scheme begins by assigning the first observation to a new category $c = 1$, and cycles through further observations, in order, assigning each to a category c according to eq. (12), updating the estimates $\hat{\mu}_c$ and σ_c^2 for a given category's mean and variance as and when new observations are assigned. Equivalent updating schemes have been described in the machine-learning literature [12], and have been proposed for models of broader human categorisation behaviour [1, 11]. By deterministically choosing the MAP estimate for category assignment at each new data-point, the scheme allows extremely rapid inference, produces a single set of category assignments, and captures order-effects that are known to be characteristic of human cognition.

Figure 1 visualises an example of the model's inferences given noisy data generated from potentially overlapping phonetic categories. The figure shows the sets of categories inferred by the model, under four differing values for α, after observing $N = 30$ speech sounds sampled randomly with equal probability from three phonetic categories with parameters $\mu_1 = 0.2$, $\sigma_1^2 = 0.01$, $\mu_2 = 0.5$, $\sigma_2^2 = 0.01$, and $\mu_3 = 0.8$, $\sigma_3^2 = 0.01$. As is clear, in an individual

Figure 1: Categories inferred by the model given $N = 30$ observed speech sounds (shown as stars on the dotted horizontal line) generated randomly from three underlying phonetic categories (top, grey), for four values of α. Prior parameters $\Lambda = (\mu_0 = 1., k_0 = 0.01 \sigma_0^2 = 0.01, \nu_0 = 2.)$ encode a vague prior that is essentially uniform across means, but weakly favours lower variance for individual categories.



learning scenario, the bias to infer more or fewer underlying categories (α) has a considerable impact on the system inferred. At the extremes: higher values (e.g. $\alpha = 5.$, second row from the top, in red) cause the model to over-fit the data by inferring 5 distinct categories; lower values (e.g. $\alpha = 0.01$, bottom row, in blue) cause the model to discount variation in the data and hypothesise a single underlying category responsible for all observations.

3. RESULTS: CULTURAL TRANSMISSION OF PHONETIC CATEGORIES

Here I implement the model of phonetic category learning in a simple simulation of cultural transmission along a chain of learners. I assume categories are transmitted via *iterated learning* [4]: each learner observes a set of (unlabelled) speech sounds X^{t-1} generated from the phonetic categories of a previous learner, induces its own set of categories $\mathbf{Z}^{*\mathbf{t}}$, then uses these inferred categories to produce speech sounds X^t that form the observations of the next learner in the chain. The population-level characteristics of this process can be obtained by averaging over inferences made by learners along the full length of the chain.

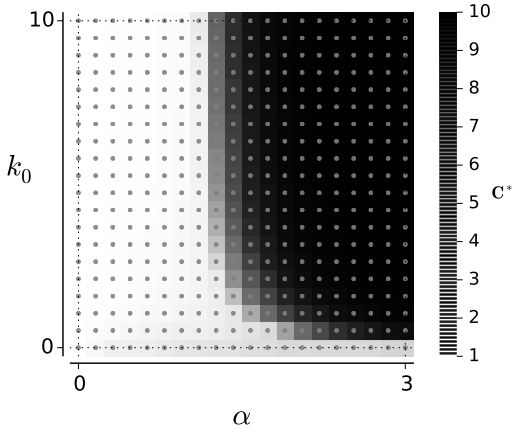
The hypothesis I set out in section 1 – that cultural transmission can be said to exhibit *universality* with respect to characteristics of individual biases – can be tested by exploring how these population-level characteristics vary – or not – as a function of the individual biases of learners in the chain. For example, here I focus on the average *number* of phonetic categories induced by learners along the chain, and how this varies with respect to the nature of two inductive biases: the concentration parameter α which dictates a bias to hypothesise more ($\alpha \to \infty$) or fewer ($\alpha \to 0$) distinct categories; and k_0, the *strength* of learners' bias in favour of particular values for the phonetic feature f. In particular, I analyse the model of transmission to ask how how these biases influence the likelihood of a phonemic merger: a collapsing of multiple distinct phonetic categories into one.

Each simulation was initialised with two distinct but overlapping phonetic categories ($\mu_1 = 0.5, \sigma_1 = 0.1, \mu_2 = 0.8, \sigma_2 = 0.1$), and run for 500 *generations*, or transmission episodes. Each learner observed $N = 10$ speech sounds generated randomly from the phonetic categories induced by the previous learner. All learners in the population share a bias to expect low variance in individual categories ($\sigma^0 = .01, \nu_0 = 10$), and a preference toward a particular range of speech sounds (an arbitrary value of f - this could represent a perceptual bias, for example), implemented by setting $\mu_0 = 0.6$. The strength of this preference for particular sounds is determined by the parameter $k_0 > 0$ (higher k_0 = stronger bias).

Figure 2 shows how the population-level average, $\mathbf{C}^*$, of the *number* of phonetic categories induced by learners varies as a function of k_0 and α. Each point in the grid represents a particular (k_0, α) pair, and the shade of the surrounding square gives $\mathbf{C}^*$, averaged over five replications of each simulation. Since the simulations were initialised with 2 phonetic categories, lighter shades ($\mathbf{C}^* < 2$) correspond to a phonetic merger, while darker shades ($\mathbf{C}^* > 2$) correspond to a *split*, or an increase in the number of categories.

While differences in α led to noticeably different inferences in *individual* learning (see figure 1), fine differences in α do not lead to fine differences in $\mathbf{C}^*$ over the course of cultural transmission constrained by a data bottleneck ($N = 10$). $\mathbf{C}^*$ does not appear to vary with the *strength* of α, only being sensitive to whether it is above or below a critical value. Roughly, any value of $\alpha < 1.5$ leads to a merger,

Figure 2: The population level average, $\mathbf{C}^*$, of the number of phonetic categories C induced by learners, as a function of individual biases α and k_0. $N = 10$, $\mu_0 = 0.6$, $\sigma_0^2 = 0.01$, $v_0 = 10$.



while $\alpha > 1.5$ causes splits. While k_0 *can* play a role in determining $\mathbf{C}^*$ (there are some non-uniform columns in the grid), it does not appear to do so under large regions of the parameter space. This insensitivity results from the modifications I made to the iMOG to allow a psychologically lightweight scheme for inference: MAP learning is known to lead to bias amplification over cultural transmission [4], and here brings about wide-scale bias-strength insensitivity.

4. DISCUSSION: TRANSMISSION, UNIVERSALITY, AND PHONETIC SYSTEMS

It is an intriguing possibility in its own right that, in at least some respects, cultural transmission of language can be likened to a broad class of physical systems via the concept of *universality*. The study of language transmission, and culture in general, may be enriched by exploring these connections further, in search of existing results relevant to dynamical systems that exhibit this kind of behaviour.

More specifically, universality in cultural transmission may have non-trivial consequences for understanding the origins of phonetic systems. For instance, if cultural transmission shapes language to match our phonetic biases, then universality implies that the process has more ammunition to work with: for example, pre-existing or domain-independent biases, even if extremely weak, could nevertheless be harnessed to shape phonetic systems.

This principle of insensitivity also implies an interesting asymmetry in our understanding of the relationship between phonetic biases and linguistic structure: for example, on the one hand, if we know a phonetic bias exists but do not know how significant a constraint it imposes, we could nevertheless make strong predictions about its population-level influence through cultural evolution; however, on the other hand, given only the population-level distributions of categories shown in figure 2, we could *not* make straightforward, reliable inferences about the *strength* of the underlying phonetic biases.

In general, wherever culture exhibits universality, we should be cautious of making direct inferences about cognition from the distributions of sounds we observe in a language.

5. SUMMARY

In this paper, I suggested that the concept of universality may be a useful – and in principle plausible – way to understand aspects of cultural transmission and the origins of our phonetic capacities. I took a recent, powerful model of unsupervised phonetic category induction, and reformulated some crucial assumptions to reflect a psychologically lightweight model of approximate inference. I simulated cultural transmission of phonetic categories under this revised model, and found conditions that suggest a dynamic reminiscent of universality in physical and multi-agent systems. While the results of these analyses reflect a very specific model, I hope to have demonstrated that the analogy is worthy of further investigation through more general means.

6. REFERENCES

[1] Anderson, J. R. 1991. The adaptive nature of human categorization. *Psychological Review* 98(3), 409–429.

[2] de Boer, B., Kuhl, P. K. Aug. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online* 4(4), 129.

[3] Feldman, N. H., Griffiths, T. L., Goldwater, S., Morgan, J. L. Oct. 2013. A role for the developing lexicon in phonetic category acquisition. *Psychological review* 120(4), 751–78.

[4] Kirby, S., Dowman, M., Griffiths, T. L. 2007. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences* 104(12), 5241–5245.

[5] Krbálek, M., Seba, P. July 2000. The statistical properties of the city transport in Cuernavaca (Mexico) and random matrix ensembles. *Journal of Physics A: Mathematical and General* 33(26), L229–L234.

[6] McMurray, B., Aslin, R. N., Toscano, J. C. Apr. 2009. Statistical learning of phonetic categories: insights from a computational approach. *Develop-*

mental science 12(3), 369–78.

[7] Montgomery, H. L. 1973. The pair correlation of zeros of the zeta function. *Analytic Number Theory, Proceedings of the 1972 St. Louis Symposiu*. American Mathematical Society 181–193.

[8] Pajak, B., Bicknell, K., Levy, R. 2013. No A model of generalization in distributional learning of phonetic categories. *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics* Sofia, Bulgaria.

[9] Parunak, H. V. D., Brueckner, S., Savit, R. 2004. Universality in Multi-Agent Systems. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems* New York, New York, USA. 930–937.

[10] Perfors, A., Tenenbaum, J. B., Regier, T. Mar. 2011. The learnability of abstract syntactic principles. *Cognition* 118(3), 306–38.

[11] Sanborn, A. N., Griffiths, T. L., Navarro, D. J. Oct. 2010. Rational approximations to rational models: alternative algorithms for category learning. *Psychological review* 117(4), 1144–67.

[12] Wang, L., Dunson, D. B. Jan. 2011. Fast Bayesian Inference in Dirichlet Process Mixture Models. *Journal of computational and graphical statistics* 20(1).