A Deep Generative Model of Vowel Formant Typology

Ryan Cotterell and **Jason Eisner**

Department of Computer Science
Johns Hopkins University, Baltimore MD, 21218
{ryan.cotterell,eisner}@jhu.edu

Abstract

What makes some types of languages more probable than others? For instance, we know that almost all spoken languages contain the vowel phoneme /i/; why should that be? The field of linguistic typology seeks to answer these questions and, thereby, divine the mechanisms that underlie human language. In our work, we tackle the problem of vowel system typology, i.e., we propose a generative probability model of which vowels a language contains. In contrast to previous work, we work directly with the acoustic information—the first two formant values—rather than modeling discrete sets of phonemic symbols (IPA). We develop a novel generative probability model and report results based on a corpus of 233 languages.

1 Introduction

Human languages are far from arbitrary; cross-linguistically, they exhibit surprising similarity in many respects and many properties appear to be universally true. The field of linguistic typology seeks to investigate, describe and quantify the axes along which languages vary. One facet of language that has been the subject of heavy investigation is the nature of vowel inventories, i.e., which vowels a language contains. It is a cross-linguistic universal that all spoken languages have vowels (Gordon, 2016), and the underlying principles guiding vowel selection are understood: vowels must be both easily recognizable and well-dispersed (Schwartz et al., 2005). In this work, we offer a more formal treatment of the subject, deriving a generative probability model of vowel inventory typology. Our work builds on (Cotterell and Eisner, 2017) by investigating not just discrete IPA inventories but the cross-linguistic variation in acoustic formants.

The philosophy behind our approach is that linguistic typology should be treated probabilistically and its goal should be the construction of a universal prior over potential languages. A probabilistic approach does not rule out linguistic systems completely (as long as one's theoretical formalism can describe them at all), but it can position phenomena on a scale from very common to very improbable. Probabilistic modeling also provides a discipline for drawing conclusions from sparse data. While we know of over 7000 human languages, we have some sort of linguistic analysis for only 2300 of them (Comrie et al., 2013), and the dataset used in this paper (Becker-Kristal, 2010) provides simple vowel data for fewer than 250 languages.

Formants are the resonant frequencies of the human vocal tract during the production of speech sounds. We propose a Bayesian generative model of vowel inventories, where each language's inventory is a finite subset of acoustic vowels represented as points $(F_1, F_2) \in \mathbb{R}^2$. We deploy tools from the neural-network and point-process literatures and experiment on a dataset with 233 distinct languages. We show that our most complicated model outperforms simpler models.

2 Acoustic Phonetics and Formants

Much of human communication takes place through speech: one conversant emits a sound wave to be comprehended by a second. In this work, we consider the nature of the portions of such sound waves that correspond to vowels. We briefly review the relevant bits of acoustic phonetics so as to give an overview of the data we are actually modeling and develop our notation.

The anatomy of a sound wave. The sound wave that carries spoken language is a function from time to amplitude, describing sound pressure variation in the air. To distinguish vowels, it is helpful to transform this function into a **spectrogram** (Fig. 1) by using a short-time Fourier transform

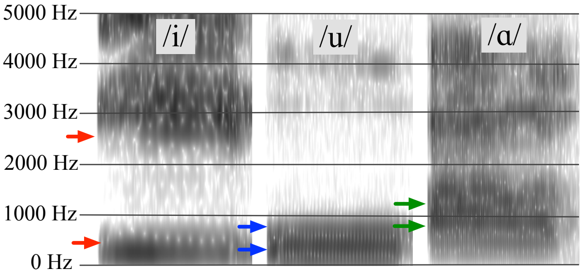

Figure 1: Example spectrogram of the three English vowels: /i/, /u/ and /ɑ/. The x-axis is time and y-axis is frequency. The first two formants F_1 and F_2 are marked in with arrows for each vowel. The figure was made with Praat (Boersma et al., 2002).

(Deng and O'Shaughnessy, 2003, Chapter 1) to decompose each short interval of the wave function into a weighted sum of sinusoidal waves of different frequencies (measured in Hz). At each interval, the variable darkness of the spectrogram indicates the weights of the different frequencies. In phonetic analysis, a common quantity to consider is a **formant**—a local maximum of the (smoothed) frequency spectrum. The fundamental frequency F_0 determines the pitch of the sound. The formants F_1 and F_2 determine the quality of the vowel.

Two is all you need (and what we left out). In terms of vowel recognition, it is widely speculated that humans rely almost exclusively on the first two formants of the sound wave (Ladefoged, 2001, Chapter 5). The two-formant assumption breaks down in edge cases: e.g., the third formant F_3 helps to distinguish the roundness of the vowel (Ladefoged, 2001, Chapter 5). Other non-formant features may also play a role. For example, in tonal languages, the same vowel may be realized with different tones (which are signaled using F_0): Mandarin Chinese makes a distinction between mǎ (*horse*) and má (*hemp*) without modifying the quality of the vowel /a/. Other features, such as creaky voice, can play a role in distinguishing phonemes. We do not explicitly model any of these aspects of vowel space, limiting ourselves to (F_1, F_2) as in previous work (Liljencrants and Lindblom, 1972). However, it would be easy to extend all the models we will propose here to incorporate such information, given appropriate datasets.

3 The Phonology of Vowel Systems

The vowel inventories of the world's languages display clear structure and appear to obey several underlying principles. The most prevalent of these principles are **focalization** and **dispersion**.

Focalization. The notion of focalization grew out of quantal vowel theory (Stevens, 1989). Quantal vowels are those that are phonetically "better" than others. They tend to display certain properties, e.g., the formants tend to be closer together (Stevens, 1987). Cross-linguistically, quantal vowels are the most frequently attested vowels, e.g., the cross-linguistically common vowel /i/ is considered quantal, but less common /y/ is not.

Dispersion. The second core principle of vowel system organization is known as dispersion. As the name would imply, the principle states that the vowels in "good" vowel systems tend to be spread out. The motivation for such a principle is clear—a well-dispersed set of vowels reduces a listener's potential confusion over which vowel is being pronounced. See Schwartz et al. (1997) for a review of dispersion in vowel system typology and its interaction with focalization, which has led to the joint dispersion-focalization theory.

Notation. We will denote the universal set of international phonetic alphabet (IPA) symbols as $\mathcal{V}$. The observed vowel inventory for language ℓ has size n^ℓ and is denoted $V^\ell = \{(v_1^\ell, \mathbf{v}_1^\ell), \ldots, (v_{n^\ell}^\ell, \mathbf{v}_{n^\ell}^\ell)\} \subseteq \mathcal{V} \times \mathbb{R}^d$, where for each $k \in [1, n^\ell]$, $v_k^\ell \in \mathcal{V}$ is an IPA symbol assigned by a linguist and $\mathbf{v}_k^\ell \in \mathbb{R}^d$ is a vector of d measurable phonetic quantities. In short, the IPA symbol v_k^ℓ was assigned as a label for a phoneme with pronunciation $\mathbf{v}_k^\ell$. The ordering of the elements within V^ℓ is arbitrary.

Goals. This framework recognizes that the same IPA symbol v (such as /u/) may represent a slightly different sound $\mathbf{v}$ in one language than in another, although they are transcribed identically. We are specifically interested in how the vowels in a language influence one another's fine-grained pronunciation in $\mathbb{R}^d$. In general, there is no reason to suspect that speakers of two languages, whose phonological systems contain the same IPA symbol, should produce that vowel with identical formants.

Data. For the remainder of the paper, we will take $d = 2$ so that each $\mathbf{v} = (F_1, F_2) \in \mathbb{R}^2$, the vector consisting of the first two formant values, as compiled from the field literature by Becker-Kristal (2006). This dataset provides inventories V^ℓ in the form above. Thus, we do not consider further variation of the vowel pronunciation that may occur within the language (between speakers, between tokens of the vowel, or between earlier and later intervals within a token).

4 Phonemes versus Phones

Previous work (Cotterell and Eisner, 2017) has placed a distribution over discrete phonemes, ignoring the variation across languages in the *pronunciation* of each phoneme. In this paper, we crack open the phoneme abstraction, moving to a learned set of finer-grained phones.

Cotterell and Eisner (2017) proposed (among other options) using a *determinantal point process* (DPP) over a universal inventory $\mathcal{V}$ of 53 symbolic (IPA) vowels. A draw from such a DPP is a language-specific inventory of vowel *phonemes*, $V \subseteq \mathcal{V}$. In this paper, we say that a language instead draws its inventory from a larger set $\bar{\mathcal{V}}$, again using a DPP. In both cases, the reason to use a DPP is that it prefers relatively diverse inventories whose individual elements are relatively quantal.

While we could in principle identify $\bar{\mathcal{V}}$ with $\mathbb{R}^d$, for convenience we still take it to be a (large) discrete finite set $\bar{\mathcal{V}} = \{\bar{v}_1, \ldots, \bar{v}_N\}$, whose elements we call *phones*. $\bar{\mathcal{V}}$ is a learned cross-linguistic parameter of our model; thus, its elements—the "universal phones"—may or may not correspond to phonetic categories traditionally used by linguists.

We presume that language ℓ draws from the DPP a subset $\bar{V}^\ell \subseteq \bar{\mathcal{V}}$, whose size we call n^ℓ. For each universal phone $\bar{v}_i$ that appears in this inventory $\bar{V}^\ell$, the language then draws an observable language-specific pronunciation $\mathbf{v}_i^\ell \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \sigma^2 I\right)$ from a distribution associated cross-linguistically with the universal phone $\bar{v}_i$. We now have an inventory of pronunciations.

As a final step in generating the vowel inventory, we could model IPA labels. For each $\bar{v}_i \in \bar{V}^\ell$, a field linguist presumably draws the IPA label v_i^ℓ conditioned on all the pronunciations $\{\mathbf{v}_i^\ell \in \mathbb{R}^d : \bar{v}_i \in \bar{V}^\ell\}$ in the inventory (and perhaps also on their underlying phones $\bar{v}_i \in \bar{V}^\ell$). This labeling process may be complex. While each pronunciation in $\mathbb{R}^d$ (or each underlying phone in $\bar{\mathcal{V}}$) may have a preference for certain IPA labels in $\mathcal{V}$, the n^ℓ labels must be drawn jointly because the linguist will take care not to use the same label for two phones, and also because the linguist may like to describe the inventory using a small number of distinct IPA features, which will tend to favor factorial grids of symbols. The linguist's use of IPA features may also be informed by phonological and phonetic processes in the language. We leave modeling of this step to future work; so our current likelihood term ignores the evidence contributed by the IPA labels in the dataset, considering only the pronunciations in $\mathbb{R}^d$.

The overall idea is that human languages ℓ draw their inventories from some universal prior, which we are attempting to reconstruct. A caveat is that we will train our method by maximum-likelihood, which does not quantify our uncertainty about the reconstructed parameters. An additional caveat is that some languages in our dataset are related to one another, which belies the idea that they were drawn independently. Ideally, one ought to capture these relationships using hierarchical or evolutionary modeling techniques.

5 Determinantal Point Processes

Before delving into our generative model, we briefly review technical background used by Cotterell and Eisner (2017). A DPP is a probability distribution over the subsets of a *fixed ground set* of size N—in our case, the set of phones $\bar{\mathcal{V}}$. The DPP is usually given as an L-ensemble (Borodin and Rains, 2005), meaning that it is parameterized by a positive semi-definite matrix $L \in \mathbb{R}^{N \times N}$. Given a discrete base set $\bar{\mathcal{V}}$ of phones, the probability of a subset $\bar{V} \subseteq \bar{\mathcal{V}}$ is given by

$$p(\bar{V}) \propto \det\left(L_{\bar{V}}\right), \qquad (1)$$

where $L_{\bar{V}}$ is the submatrix of L corresponding to the rows and columns associated with the subset $\bar{V} \subseteq \bar{\mathcal{V}}$. The entry L_{ij}, where $i \neq j$, has the effect of describing the similarity between the elements $\bar{v}_i$ and $\bar{v}_j$ (both in $\bar{\mathcal{V}}$)—an ingredient needed to model dispersion. And, the entry L_{ii} describes the quality—focalization—of the vowel $\bar{v}_i$, i.e., how much the model wants to have $\bar{v}_i$ in a sampled set independent of the other members.

5.1 Probability Kernel

In this work, each phone $\bar{v}_i \in \bar{\mathcal{V}}$ is associated with a probability density over the space of possible pronunciations $\mathbb{R}^2$. Our measure of phone similarity will consider the "overlap" between the densities associated with two phones. This works as follows: Given two densities $f(x, y)$ and $f'(x, y)$ over $\mathbb{R}^2$, we define the kernel (Jebara et al., 2004) as

$$\mathcal{K}(f, f'; \rho) = \int_x \int_y f(x, y)^\rho f'(x, y)^\rho dx\, dy, \quad (3)$$

$$\prod_{\ell=1}^{M} \Big[p(\mathbf{v}^{\ell,1}, \ldots, \mathbf{v}^{\ell,n^\ell} \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, N) \Big] p(\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_N \mid N)\, p(N) \tag{2}$$

$$= \prod_{\ell=1}^{M} \left[\sum_{\mathbf{a}^\ell \in A(n^\ell, N)} \left(\prod_{k=1}^{n^\ell} \underbrace{p(\mathbf{v}^{\ell,k} \mid \boldsymbol{\mu}_{a_k^\ell})}_{\textcircled{4}} \right) \underbrace{p(\bar{V}(\mathbf{a}^\ell) \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, N)}_{\textcircled{3}} \right] \underbrace{p(\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_N \mid N)}_{\textcircled{2}} \underbrace{p(N)}_{\textcircled{1}}$$

Figure 2: Joint likelihood of M vowel systems under our deep generative probability model for continuous-space vowel inventories. Here language ℓ has an observed inventory of pronunciations $\{\mathbf{v}^{\ell,k} : 1 \leq k \leq n^\ell\}$, and $a_k^\ell \in [1, N]$ denotes a phone that might be responsible for the pronunciation $\mathbf{v}^{\ell,k}$. Thus, $\mathbf{a}^\ell$ denotes some way to jointly label all n^ℓ pronunciations with distinct phones. We must sum over all $\binom{N}{n^\ell}$ such labelings $\mathbf{a}^\ell \in A(n^\ell, N)$ since the true labeling is not observed. In other words, we sum over all ways $\mathbf{a}^\ell$ of completing the data for language ℓ. Within each summand, the product of factors 3 and 4 is the probability of the completed data, i.e., the joint probability of generating the inventory $\bar{V}(\mathbf{a}^\ell)$ of phones used in the labeling and their associated pronunciations. Factor 3 considers the prior probability of $\bar{V}(\mathbf{a}^\ell)$ under the DPP, and factor 4 is a likelihood term that considers the probability of the associated pronunciations.

with inverse temperature parameter ρ.

In our setting, f, f' will both be Gaussian distributions with means $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ that share a fixed spherical covariance matrix $\sigma^2 I$. Then eq. (3) and indeed its generalization to any $\mathbb{R}^d$ has a closed-form solution (Jebara et al., 2004, §3.1):

$$\mathcal{K}(f, f'; \rho) = \tag{4}$$
$$(2\rho)^{\frac{d}{2}} \left(2\pi\sigma^2\right)^{\frac{(1-2\rho)d}{2}} \exp\left(-\frac{\rho ||\boldsymbol{\mu} - \boldsymbol{\mu}'||^2}{4\sigma^2} \right).$$

Notice that making ρ small (i.e., high temperature) has an effect on (4) similar to scaling the variance σ^2 by the temperature, but it also results in changing the scale of $\mathcal{K}$, which affects the balance between dispersion and focalization in (6) below.

5.2 Focalization Score

The probability kernel given in eq. (3) naturally handles the linguistic notion of dispersion. What about focalization? We say that a phone is focal to the extent that it has a high score

$$F(\boldsymbol{\mu}) = \exp\left(U_2 \tanh(U_1 \boldsymbol{\mu} + \mathbf{b}_1) + \mathbf{b}_2\right) > 0 \tag{5}$$

where $\boldsymbol{\mu}$ is the mean of its density. To learn the parameters of this neural network from data is to learn which phones are focal. We use a neural network since the focal regions of $\mathbb{R}^2$ are distributed in a complex way.

5.3 The L Matrix

Algorithm 1 Generative Process

```
$N \sim \text{Poisson}\,(\lambda)\ (\in \mathbb{N})$   ①
for $i = 1$ to $N$ :
    $\boldsymbol{\mu}_i \sim \mathcal{N}\left(\mathbf{0}, I\right)\ (\in \mathbb{R}^2)$   ②
define $L \in \mathbb{R}^{N \times N}$ via (6)
for $\ell = 1$ to $M$ :
    $\bar{V}^\ell \sim \text{DPP}\,(L)\ (\subseteq [1, N])$; let $n^\ell = |\bar{V}^\ell|$   ③
    for $i \in \bar{V}^\ell$ :
        $\tilde{\mathbf{v}}_i^\ell \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \sigma^2 I\right)$   ④
        $\mathbf{v}_i^\ell = \nu_{\boldsymbol{\theta}}\left(\tilde{\mathbf{v}}_i^\ell\right)$   ④
```

If $f_i = \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 I)$ is the density associated with the phone $\bar{v}_i$, we may populate an $N \times N$ real matrix L where

$$L_{ij} = \begin{cases} \mathcal{K}(f_i, f_j; \rho) & \text{if } i \neq j \\ \mathcal{K}(f_i, f_j; \rho) + F(\boldsymbol{\mu}_i) & \text{if } i = j \end{cases} \tag{6}$$

Since L is the sum of two positive definite matrices (the first specializes a known kernel and the second is diagonal and positive), it is also positive definite. As a result, it can be used to parameterize a DPP over $\bar{\mathcal{V}}$. Indeed, since L is positive definite and not merely positive semidefinite, it will assign positive probability to *any* subset of $\bar{\mathcal{V}}$.

As previously noted, this DPP does not define a distribution over an infinite set, e.g., the powerset of $\mathbb{R}^2$, as does recent work on continuous DPPs (Affandi et al., 2013). Rather, it defines a distribution over the powerset of a *set of densities with finite cardinality*. Once we have sampled a subset of densities, a real-valued quantity may be additionally sampled from each sampled density.

6 A Deep Generative Model

We are now in a position to expound our generative model of continuous-space vowel typology. We generate a set of formant pairs for M languages in a four step process. Note that throughout this exposition, language-specific quantities with be superscripted with an integral language marker ℓ, whereas universal quantities are left unsuperscripted. The generative process is written in algorithmic form in Alg. 1. Note that each step is numbered and color-coded for ease of comparison with the full joint likelihood in Fig. 2.

Step ①: $p(N)$**.** We sample the size N of the universal phone inventory $\bar{\mathcal{V}}$ from a Poisson distribution with a rate parameter λ, i.e.,

$$N \sim \text{Poisson}\left(\lambda\right). \tag{7}$$

That is, we do not presuppose a certain number of phones in the model.

Step ②: $p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N)$**.** Next, we sample the means $\boldsymbol{\mu}_i$ of the Gaussian phones. In the model presented here, we assume that each phone is generated independently, so $p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N) = \prod_{i=1}^{N} p(\boldsymbol{\mu}_i)$. Also, we assume a standard Gaussian prior over the means, $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, I)$.

The sampled means define our N Gaussian phones $\mathcal{N}\left(\boldsymbol{\mu}_i, \sigma^2 I\right)$: we are assuming for simplicity that all phones share a *single* spherical covariance matrix, defined by the hyperparameter σ^2. The dispersion and focalization of these phones define the matrix L according to equations (4)–(6), where ρ in (4) and the weights of the focalization neural net (5) are also hyperparameters.

Step ③: $p(\bar{V}^\ell \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N)$**.** Next, for each language $\ell \in [1, \ldots, M]$, we sample a diverse subset of the N phones, via a single draw from a DPP parameterized by matrix L:

$$\bar{V}^\ell \sim \text{DPP}(L), \tag{8}$$

where $\bar{V}^\ell \subseteq [1, N]$. Thus, $i \in \bar{V}^\ell$ means that language ℓ contains phone $\bar{v}_i$. Note that even the size of the inventory, $n^\ell = |\bar{V}^\ell|$, was chosen by the DPP. In general, we have $n^\ell \ll N$.

Step ④: $\prod_{i \in \bar{V}^\ell} p(\mathbf{v}_i^\ell \mid \boldsymbol{\mu}_i)$ The final step in our generative process is that the phones $\bar{v}_i$ in language ℓ must generate the pronunciations $\mathbf{v}_i^\ell \in \mathbb{R}^2$ (formant vectors) that are actually observed in language ℓ. Each vector takes two steps. For each $i \in \bar{V}^\ell$, we generate an underlying $\tilde{\mathbf{v}}_i \in \mathbb{R}^2$ from the corresponding Gaussian phone. Then, we run this vector through a feed-forward neural network $\nu_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$. In short:

$$\tilde{\mathbf{v}}_i^\ell \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 I) \tag{9}$$

$$\mathbf{v}_i^\ell = \nu_{\boldsymbol{\theta}}(\tilde{\mathbf{v}}_i^\ell), \tag{10}$$

where the second step is deterministic. We can fuse these two steps into a single step $p(\mathbf{v}_i \mid \boldsymbol{\mu}_i)$, whose closed-form density is given in eq. (12) below. In effect, step 4 takes a Gaussian phone as input and produces the observed formant vector with an *underlying* formant vector in the middle.

This completes our generative process. We do not observe all the steps, but only the final collection of pronunciations $\mathbf{v}_i^\ell$ for each language, where the subscripts i that indicate phone identity have been lost. The probability of this incomplete dataset involves summing over possible phones for each pronunciation, and is presented in Fig. 2.

6.1 A Neural Transformation of a Gaussian

A crucial bit of our model is running a sample from a Gaussian through a neural network. Under certain restrictions, we can find a closed form for the resulting density; we discuss these below. Let $\nu_{\boldsymbol{\theta}}$ be a depth-2 multi-layer perceptron

$$\nu_{\boldsymbol{\theta}}(\tilde{\mathbf{v}}_i) = W_2 \tanh\left(W_1 \tilde{\mathbf{v}}_i + \mathbf{b}_1\right) + \mathbf{b}_2. \tag{11}$$

In order to find a closed-form solution, we require that (5) be a diffeomorphism, i.e., an invertible mapping from $\mathbb{R}^2 \to \mathbb{R}^2$ where both $\nu_{\boldsymbol{\theta}}$ and its inverse $\nu_{\boldsymbol{\theta}}^{-1}$ are differentiable. This will be true as long as $W_1, W_2 \in \mathbb{R}^{2\times 2}$ are square matrices of full-rank and we choose a smooth, invertible activation function, such as $\tanh$. Under those conditions, we may apply the standard theorem for transforming a random variable (see Stark and Woods, 2011):

$$\begin{aligned} p(\mathbf{v}_i \mid \boldsymbol{\mu}_i) &= p(\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{v}_i) \mid \boldsymbol{\mu}_i) \det J_{\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{v}_i)} \\ &= p(\tilde{\mathbf{v}}_i \mid \boldsymbol{\mu}_i) \det J_{\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{v}_i)} \end{aligned} \tag{12}$$

where $J_{\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{x})}$ is the Jacobian of the inverse of the neural network at the point $\mathbf{x}$. Recall that $p(\tilde{\mathbf{v}}_i \mid \boldsymbol{\mu}_i)$ is Gaussian-distributed.

7 Modeling Assumptions

Imbued in our generative story are a number of assumptions about the linguistic processes behind vowel inventories. We briefly draw connections between our theory and the linguistics literature.

Why underlying phones? A technical assumption of our model is the existence of a universal set of underlying phones. Each phone is equipped with a probability distribution over reported acoustic measurements (pronunciations), to allow for a single phone to account for multiple slightly different pronunciations in different languages (though never in the same language). This distribution can capture both actual interlingual variation and also random noise in the measurement process.

While our universal phones may seem to resemble the universal IPA symbols used in phonological transcription, they lack the rich featural specifications of such phonemes. A phone in our model has no features other than its mean position, which wholly determines its behavior. Our universal phones are not a substantive linguistic hypothesis, but are essentially just a way of partitioning $\mathbb{R}^2$ into finitely many small regions whose similarity and focalization can be precomputed. This technical trick allows us to use a discrete rather than a continuous DPP over the $\mathbb{R}^2$ space.[1]

Why a neural network? Our phones are Gaussians of spherical variance σ^2, presumed to be scattered with variance 1 about a two-dimensional *latent* vowel space. Distances in this latent space are used to compute the dissimilarity of phones for modeling dispersion, and also to describe the phone's ability to vary across languages. That is, two phones that are *distant* in the latent space can appear in the same inventory—presumably they are easy to discriminate in both perception and articulation—and it is easy to choose which one better explains an acoustic measurement, thereby affecting the other measurements that may appear in the inventory.

We relate this *latent* space to measurable acoustic space by a learned diffeomorphism $\nu_{\boldsymbol{\theta}}$ (Cotterell and Eisner, 2017). $\nu_{\boldsymbol{\theta}}^{-1}$ can be regarded as warping the acoustic distances into perceptual/articulatory distances. In some "high-resolution" regions of acoustic space, phones with fairly similar (F_1, F_2) values might yet be far apart in the latent space. Conversely, in other regions, relatively large acoustic changes in some direction might not prevent two phones from acting as similar or two pronunciations from being attributed to the same phone. In general, a unit circle of radius σ in latent space may be mapped by $\nu_{\boldsymbol{\theta}}$ to an oddly shaped connected region in acoustic space, and a Gaussian in latent space may be mapped to a multimodal distribution.

8 Inference and Learning

We fit our model via MAP-EM (Dempster et al., 1977). The E-step involves deciding which phones each language has. To achieve this, we fashion a Gibbs sampler (Geman and Geman, 1984), yielding a Markov-Chain Monte Carlo E-step (Levine and Casella, 2001).

8.1 Inference: MCMC E-Step

Inference in our model is intractable even when the phones $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ are fixed. Given a language with n vowels, we have to determine which subset of the N phones best explains those vowels. As discussed above, the alignment $\mathbf{a}$ between the n vowels and n of the N phones represents a latent variable. Marginalizing it out is #P-hard, as we can see that it is equivalent to summing over all bipartite matchings in a weighted graph, which, in turn, is as costly as computing the permanent of a matrix (Valiant, 1979). Our sampler[2] is an approximation algorithm for the task. We are interested in sampling $\mathbf{a}$, the labeling of observed vowels with universal phones. Note that this implicitly samples the language's phone inventory $\bar{V}(\mathbf{a})$, which is fully determined by $\mathbf{a}$.

Specifically, we employ an MCMC method closely related to Gibbs sampling. At each step of the sampler, we update our vowel-phone alignment $\mathbf{a}^\ell$ as follows. Choose a language ℓ and a vowel index $k \in [1, n^\ell]$, and let $i = a_k^\ell$ (that is, pronunciation $\mathbf{v}^{\ell,k}$ is currently labeled with universal phone $\bar{v}_i$). We will consider changing a_k^ℓ to j, where j is drawn from the $(N - n^\ell)$ phones that do *not* appear in $\bar{V}(\mathbf{a}^\ell)$, heuristically choosing j in proportion to the likelihood $p(\mathbf{v}^{\ell,k} \mid \boldsymbol{\mu}_j)$. We then stochastically decide whether to keep $a_k^\ell = i$ or set $a_k^\ell = j$ in proportion to the resulting values of the product ④ · ③ in eq. (2).

For a single E-step, the Gibbs sampler "warm-starts" with the labeling from the end of the previous iteration's E-step. It sweeps $S = 5$ times

[1] Indeed, we could have simply taken our universal phone set to be a huge set of tiny, regularly spaced overlapping Gaussians that "covered" (say) the unit circle. As a computational matter, we instead opted to use a smaller set of Gaussians, giving the learner the freedom to infer their positions and tune their variance σ^2. Because of this freedom, this set should not be too large, or a MAP learner may overfit the training data with zero-variance Gaussians and be unable to explain the test languages—similar to overfitting a Gaussian mixture model.

[2] Taken from Volkovs and Zemel (2012, 3.1).

through all vowels for all languages, and returns S sampled labelings, one from the end of each sweep.

We are also interested in automatically choosing the number of phones N, for which we take the Poisson's rate parameter $\lambda = 100$. To this end, we employ reversible-jump MCMC (Green, 1995), resampling N at the start of every E-step.

8.2 Learning: M-Step

Given the set of sampled alignments provided by the E-step, our M-step consists of optimizing the log-likelihood of the now-complete training data using the inferred latent variables. We achieved this through SGD training of the diffeomorphism parameters $\boldsymbol{\theta}$, the means $\boldsymbol{\mu}_i$ of the Gaussian phones, and the parameters of the focalization kernel $\mathcal{F}$.

9 Experiments

9.1 Data

Our data is taken from the Becker-Kristal corpus (Becker-Kristal, 2006), which is a compilation of various phonetic studies and forms the largest multilingual phonetic database. Each entry in the corpus corresponds to a linguist's phonetic description of a language's vowel system: an inventory consisting of IPA symbols where each symbol is associated with two or more formant values. The corpus contains data from 233 distinct languages. When multiple inventories were available for the same language (due to various studies in the literature), we selected one at random and discarded the others.

9.2 Baselines

Baseline #1: Removing dispersion. The key technical innovation in our work lies in the incorporation of a DPP into a generative model of vowel formants—a continuous-valued quantity. The role of the DPP was to model the linguistic principle of dispersion—we may cripple this portion of our model, e.g., by forcing $\mathcal{K}$ to be a diagonal kernel, i.e., $K_{ij} = 0$ for $i \neq j$. In this case the DPP becomes a Bernoulli Point Process (BPP)—a special case of the DPP. Since dispersion is widely accepted to be an important principle governing naturally occurring vowel systems, we expect a system trained without such knowledge to perform worse.

Baseline #2: Removing the neural network $\nu_{\boldsymbol{\theta}}$. Another question we may ask of our formulation is whether we actually need a fancy neural mapping $\nu_{\boldsymbol{\theta}}$ to model our typological data well. The human perceptual system is known to perform a non-linear transformation on acoustic signals, starting with the non-linear cochlear transform that is physically performed in the ear. While $\nu_{\boldsymbol{\theta}}^{-1}$ is intended as loosely analogous, we determine its benefit by removing eq. (10) from our generative story, i.e., we take the observed formants $\mathbf{v}_k$ to arise directly from the Gaussian phones.

Baseline #3: Supervised phones and alignments. A final baseline we consider is *supervised* phones. Linguists standardly employ a finite set of phones—symbols from the international phonetic alphabet (IPA). In phonetic annotation, it is common to map each sound in a language back to this universal discrete alphabet. Under such an annotation scheme, it is easy to discern, cross-linguistically, which vowels originate from the same phoneme: an /ɪ/ in German may be roughly equated with an /ɪ/ in English. However, it is not clear how consistent this annotation truly is. There are several reasons to expect high-variance in the cross-linguistic acoustic signal. First, IPA symbols are primarily useful for interlinked phonological distinctions, i.e., one applies the symbol /ɪ/ to distinguish it from /i/ in the given language, rather than to associate it with the sound bearing the same symbol in a second language. Second, field linguists often resort to the closest common IPA symbol, rather than an exact match: if a language makes no distinction between /i/ and /ɪ/, it is more common to denote the sound with a /i/. Thus, IPA may not be as universal as hoped. Our dataset contains 50 IPA symbols so this baseline is only reported for $N = 50$.

9.3 Evaluation

Evaluation in our setting is tricky. The scientific goal of our work is to place a bit of linguistic theory on a firm probabilistic footing, rather than a downstream engineering-task, whose performance we could measure. We consider three metrics.

Cross-Entropy. Our first evaluation metric is cross-entropy: the average negative log-probability of the vowel systems in held-out test data, given the universal inventory of N phones that we trained through EM. We find this to be the cleanest method for scientific evaluation—it is the metric of optimization and has a clear interpretation: how surprised was the model to see the vowel systems of held-out, but attested, languages?

The cross-entropy is the negative log of the $\prod \left[\cdots \right]$ expression in eq. (2), with ℓ now rang-

N	metric	DPP+$\nu_{\boldsymbol{\theta}}$	BPP+$\nu_{\boldsymbol{\theta}}$	DPP−$\nu_{\boldsymbol{\theta}}$	Sup.
	x-ent	540.02	540.05	600.34	✗
15	cloze1	5.76	5.76	6.53	✗
	cloze12	4.89	4.89	5.24	✗
	x-ent	280.47	275.36	335.36	✗
25	cloze1	5.04	5.25	6.23	✗
	cloze12	4.76	4.97	5.43	✗
	x-ent	222.85	231.70	320.05	1610.37
50	cloze1	3.38	3.16	4.02	4.96
	cloze12	2.73	2.93	3.04	6.95
	x-ent	212.14	220.42	380.31	✗
57	cloze1	2.21	3.08	3.25	✗
	cloze12	2.01	3.05	3.41	✗
	x-ent	271.95	301.45	380.02	✗
100	cloze1	2.26	2.42	3.03	✗
	cloze12	1.96	2.01	2.51	✗

Table 1: Cross-entropy in nats per language (lower is better) and expected Euclidean-distance error of the cloze prediction (lower is better). The overall best value for each task is boldfaced. The case $N = 50$ is compared against our supervised baseline. The $N = 57$ row is the case where we allowed N to fluctuate during inference using reversible-jump MCMC; this was the N value selected at the final EM iteration.

ing over held-out languages.[3] Wallach et al. (2009) give several methods for estimating the intractable sum in language ℓ. We use the simple harmonic mean estimator, based on 50 samples of $\mathbf{a}^\ell$ drawn with our Gibbs sampler (warm-started from the final E-step of training).

Cloze Evaluation. In addition, following Cotterell and Eisner (2017), we evaluate our trained model's ability to perform a cloze task (Taylor, 1953). Given $n^\ell - 1$ or $n^\ell - 2$ of the vowels in held-out language ℓ, can we predict the pronunciations $\mathbf{v}_k$ of the remaining 1 or 2? We predict $\mathbf{v}_k$ to be $\nu_{\boldsymbol{\theta}}(\boldsymbol{\mu}_i)$ where $i = a_k^\ell$ is the phone inferred by the sampler. Note that the sampler's inference here is based only on the observed vowels (the likelihood) and the focalization-dispersion preferences of the DPP (the prior). We report the expected error of such a prediction—where error is quantified by Euclidean distance in (F_1, F_2) formant space—over the same 50 samples of $\mathbf{a}^\ell$.

For instance, consider a previously unseen vowel system with formant values $\{(499, 2199), (861, 1420), (571, 1079)\}$. A "cloze1" evaluation would aim to predict $\{(499, 2199)\}$ as the missing vowel, given $\{(861, 1420), (571, 1079)\}$, and the fact that $n^\ell = 3$. A "cloze12" evaluation would aim to predict two missing vowels.

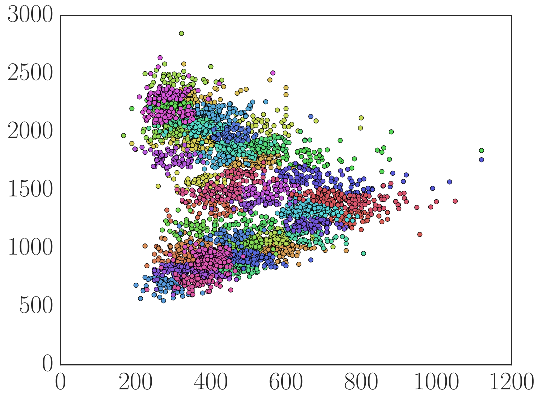

Figure 3: A graph of $\mathbf{v} = (F_1, F_2)$ in the union of all the training languages' inventories, color-coded by inferred phone ($N = 50$).

9.4 Experimental Details

Here, we report experimental details and the hyperparameters that we use to achieve the results reported. We consider a neural network $\nu_{\boldsymbol{\theta}}$ with $k \in [1, 4]$ layers and find $k = 1$ the best performer on development data. Recall that our *diffeomorphism* constraint requires that each layer have exactly two hidden units, the same as the number of observed formants. We consider $N \in \{15, 25, 50, 100\}$ phones as well as letting N fluctuate with reversible-jump MCMC (see footnote 1). We train for 100 iterations of EM, taking $S = 5$ samples at each E-step. At each M-step, we run 50 iterations of SGD for the focalization NN and also for the diffeomorphism NN. For each N, we selected (σ^2, ρ) by minimizing cross-entropy on a held-out development set. We considered $(\sigma^2, \rho) \in \{10^k\}_{k=1}^5 \times \{\rho^k\}_{k=1}^5$.

9.5 Results and Error Analysis

We report results in Tab. 1. We find that our DPP model improves over the baselines. The results support two claims: (i) dispersion plays an important role in the structure of vowel systems and (ii) learning a non-linear transformation of a Gaussian improves our ability to model sets of formant-pairs. Also, we observe that as we increase the number of phones, the role of the DPP becomes more important. We visualize a sample of the trained alignment in Fig. 3.

[3] Since that expression is the product of both probability distributions and probability densities, our "cross-entropy" metric is actually the sum of both entropy terms and (potentially negative) differential entropy terms. Thus, a value of 0 has no special significance.

Frequency Encodes Dispersion. Why does dispersion not always help? The models with fewer phones do not reap the benefits that the models with more phones do. The reason lies in the fact that the most common vowel formants are *already* dispersed. This indicates that we still have not quite modeled the mechanisms that select for good vowel formants, despite our work at the phonetic level; further research is needed. We would prefer a model that explains the *evolutionary motivation* of sound systems as communication systems.

Number of Induced Phones. What is most salient in the number of induced phones is that it is close to the number of IPA phonemes in the data. However, the performance of the phoneme-supervised system is much worse, indicating that, perhaps, while the linguists have the right idea about the *number* of universal symbols, they did not specify the correct IPA symbol in all cases. Our data analysis indicates that this is often due to pragmatic concerns in linguistic field analysis. For example, even if /ɪ/ is the proper IPA symbol for the sound, if there is no other sound in the vicinity the annotator may prefer to use more common /i/.

10 Related Work

Most closely related to our work is the classic study of Liljencrants and Lindblom (1972), who provide a simulation-based account of vowel systems. They argued that minima of a certain objective that encodes dispersion should correspond to canonical vowel systems of a given size n. Our tack is different in that we construct a generative probability model, whose parameters we learn from data. However, the essence of modeling is the same in that we explain *formant* values, rather than discrete IPA symbols. By extension, our work is also closely related to extensions of this theory (Schwartz et al., 1997; Roark, 2001) that focused on incorporating the notion of focalization into the experiments.

Our present paper can also be regarded as a continuation of Cotterell and Eisner (2017), in which we used DPPs to model vowel inventories as sets of discrete IPA symbols. That paper pretended that each IPA symbol had a single cross-linguistic (F_1, F_2) pair, an idealization that we remove in this paper by discarding the IPA symbols and modeling formant values directly.

11 Conclusion

Our model combines existing techniques of probabilistic modeling and inference to attempt to fit the actual distribution of the world's vowel systems. We presented a generative probability model of sets of measured (F_1, F_2) pairs. We view this as a necessary step in the development of generative probability models that can explain the distribution of the world's languages. Previous work on generating vowel inventories has focused on how those inventories were transcribed into IPA by field linguists, whereas we focus on the field linguists' acoustic measurements of how the vowels are actually pronounced.

Acknowledgments

We would like to acknowledge Tim Vieira, Katharina Kann, Sebastian Mielke and Chu-Cheng Lin for reading many early drafts. The first author would like to acknowledge an NDSEG grant and a Facebook PhD fellowship. This material is also based upon work supported by the National Science Foundation under Grant No. 1718846 to the last author.

References

Raja Hafiz Affandi, Emily Fox, and Ben Taskar. 2013. Approximate inference in continuous determinantal processes. In *Advances in Neural Information Processing Systems*, pages 1430–1438.

Roy Becker-Kristal. 2006. Predicting vowel inventories: The dispersion-focalization theory revisited. *The Journal of the Acoustical Society of America*, 120(5):3248–3248.

Roy Becker-Kristal. 2010. *Acoustic Typology of Vowel Inventories and Dispersion Theory: Insights from a Large Cross-Linguistic Corpus*. Ph.D. thesis, UCLA.

Paulus Petrus Gerardus Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot International*, 5.

Alexei Borodin and Eric M. Rains. 2005. Eynard-Mehta theorem, Schur process, and their Pfaffian analogs. *Journal of Statistical Physics*, 121(3-4):291–317.

Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath. 2013. Introduction. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Rtatistical Society, Series B (Statistical Methodology)*, pages 1–38.

Li Deng and Douglas O'Shaughnessy. 2003. *Speech Processing: A Dynamic and Optimization-Oriented Approach*. CRC Press.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

Matthew K. Gordon. 2016. *Phonological Typology*. Oxford.

Peter J. Green. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Tony Jebara, Risi Kondor, and Andrew Howard. 2004. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844.

Peter Ladefoged. 2001. *Vowels and Consonants: An Introduction to the Sounds of Languages*. Wiley-Blackwell.

Richard A. Levine and George Casella. 2001. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439.

Johan Liljencrants and Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, pages 839–862.

Brian Roark. 2001. Explaining vowel inventory tendencies via simulation: Finding a role for quantal locations and formant normalization. In *North East Linguistic Society*, volume 31, pages 419–434.

Jean-Luc Schwartz, Christian Abry, Louis-Jean Boë, Nathalie Vallée, and Lucie Ménard. 2005. The dispersion-focalization theory of sound systems. *The Journal of the Acoustical Society of America*, 117(4):2422–2422.

Jean-Luc Schwartz, Louis-Jean Boë, Nathalie Vallée, and Christian Abry. 1997. The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25(3):255–286.

Henry Stark and John Woods. 2011. *Probability, Statistics, and Random Processes for Engineers*. Pearson.

Kenneth N. Stevens. 1987. Relational properties as perceptual correlates of phonetic features. In *International Conference of Phonetic Sciences*, pages 352–355.

Kenneth N. Stevens. 1989. On the quantal nature of speech. *Journal of Phonetics*, 17:3–45.

Wilson L. Taylor. 1953. Cloze procedure: a new tool for measuring readability. *Journalism and Mass Communication Quarterly*, 30(4):415.

Leslie G. Valiant. 1979. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189–201.

Maksims Volkovs and Richard S. Zemel. 2012. Efficient sampling for bipartite matching problems. In *Advances in Neural Information Processing Systems*, pages 1313–1321.

Hanna Wallach, Ian Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *International Conference on Machine Learning (ICML)*, pages 1105–1112.