Maximum Likelihood Estimation of Feature-based Distributions

Jeffrey Heinz and Cesar Koirala
University of Delaware
Newark, Delaware, USA
{heinz,koirala}@udel.edu

Abstract

Motivated by recent work in phonotactic learning (Hayes and Wilson 2008, Albright 2009), this paper shows how to define feature-based probability distributions whose parameters can be provably efficiently estimated. The main idea is that these distributions are defined as a product of simpler distributions (cf. Ghahramani and Jordan 1997). One advantage of this framework is it draws attention to what is minimally necessary to describe and learn phonological feature interactions in phonotactic patterns. The "bottom-up" approach adopted here is contrasted with the "top-down" approach in Hayes and Wilson (2008), and it is argued that the bottom-up approach is more analytically transparent.

1 Introduction

The hypothesis that the atomic units of phonology are phonological features, and not segments, is one of the tenets of modern phonology (Jakobson et al., 1952; Chomsky and Halle, 1968). According to this hypothesis, segments are essentially epiphenomenal and exist only by virtue of being a shorthand description of a collection of more primitive units—the features. Incorporating this hypothesis into phonological learning models has been the focus of much influential work (Gildea and Jurafsky, 1996; Wilson, 2006; Hayes and Wilson, 2008; Moreton, 2008; Albright, 2009).

This paper makes three contributions. The first contribution is a framework within which:

1. researchers can choose which statistical independence assumptions to make regarding phonological features;
2. feature systems can be fully integrated into strictly local (McNaughton and Papert, 1971) (i.e. n-gram models (Jurafsky and Martin, 2008)) and strictly piecewise models (Rogers et al., 2009; Heinz and Rogers, 2010) in order to define families of provably well-formed, feature-based probability distributions that are provably efficiently estimable.

The main idea is to define a family of distributions as the normalized product of simpler distributions. Each simpler distribution can be represented by a Probabilistic Deterministic Finite Acceptor (PDFA), and the product of these PDFAs defines the actual distribution. When a family of distributions $\mathcal{F}$ is defined in this way, $\mathcal{F}$ may have many fewer parameters than if $\mathcal{F}$ is defined over the product PDFA directly. This is because the parameters of the distributions are defined in terms of the factors which combine in predictable ways via the product. Fewer parameters means accurate estimation occurs with less data and, relatedly, the family contains fewer distributions.

This idea is not new. It is explicit in Factorial Hidden Markov Models (FHMMs) (Ghahramani and Jordan, 1997; Saul and Jordan, 1999), and more recently underlies approaches to describing and inferring regular string transductions (Dreyer et al., 2008; Dreyer and Eisner, 2009). Although HMMs and probabilistic finite-state automata describe the same class of distributions (Vidal et al., 2005a; Vidal et al., 2005b), this paper presents these ideas in formal language-theoretic and automata-theoretic terms because (1) there are no hidden states and is thus simpler than FHMMs, (2) determinstic automata have several desirable properties crucially used here, and (3) PDFAs add probabilities to structure whereas HMMs add structure to probabilities and the authors are more comfortable with the former perspective (for further discussion, see Vidal et al. (2005a,b)).

The second contribution illustrates the main idea with a feature-based bigram model with a

Proceedings of the 11th Meeting of the ACL-SIGMORPHON, ACL 2010, pages 28–37,
Uppsala, Sweden, 15 July 2010.

strong statistical independence assumption: no two features interact. This is shown to capture exactly the intuition that sounds with like features have like distributions. Also, the assumption of non-interacting features is shown to be too strong because like sounds do not have like distributions in actual phonotactic patterns. Four kinds of featural interactions are identified and possible solutions are discussed.

Finally, we compare this proposal with Hayes and Wilson (2008). Essentially, the model here represents a "bottom-up" approach whereas theirs is "top-down." "Top-down" models, which consider every set of features as potentially interacting in every allowable context, face the difficult problem of searching a vast space and often resort to heuristic-based methods, which are difficult to analyze. To illustrate, we suggest that the role played by phonological features in the phonotactic learner in Hayes and Wilson (2008) is not well-understood. We demonstrate that classes of all segments but one (i.e. the complement classes of single segments) play a significant role, which diminishes the contribution provided by natural classes themselves (i.e. ones made by phonological features). In contrast, the proposed model here is analytically transparent.

This paper is organized as follows. §2 reviews some background. §3 discusses bigram models and §4 defines feature systems and feature-based distributions. §5 develops a model with a strong independence assumption and §6 discusses featural interaction. §7 dicusses Hayes and Wilson (2008) and §8 concludes.

2 Preliminaries

We start with mostly standard notation. $\mathcal{P}(A)$ is the powerset of A. Σ denotes a finite set of symbols and a string over Σ is a finite sequence of these symbols. Σ^+ and Σ^* denote all strings over this alphabet of nonzero but finite length, and of any finite length, respectively. A function f with domain A and codomain B is written $f : A \to B$. When discussing partial functions, the notation $\uparrow$ and $\downarrow$ indicate for particular arguments whether the function is undefined and defined, respectively.

A *language* L is a subset of Σ^*. A *stochastic language* $\mathcal{D}$ is a probability distribution over Σ^*. The probability p of word w with respect to $\mathcal{D}$ is written $Pr_{\mathcal{D}}(w) = p$. Recall that all distributions $\mathcal{D}$ must satisfy $\sum_{w \in \Sigma^*} Pr_{\mathcal{D}}(w) = 1$. If L is language then $Pr_{\mathcal{D}}(L) = \sum_{w \in L} Pr_{\mathcal{D}}(w)$. Since all distributions in this paper are stochastic languages, we use the two terms interchangeably.

A *Probabilistic Deterministic Finite-state Automaton* (PDFA) is a tuple $\mathcal{M} = \langle Q, \Sigma, q_0, \delta, F, T \rangle$ where Q is the state set, Σ is the alphabet, q_0 is the start state, δ is a deterministic transition function, F and T are the final-state and transition probabilities. In particular, $T : Q \times \Sigma \to \mathbb{R}^+$ and $F : Q \to \mathbb{R}^+$ such that

$$\text{for all } q \in Q,\ F(q) + \sum_{\sigma \in \Sigma} T(q, \sigma) = 1. \quad (1)$$

PDFAs are typically represented as labeled directed graphs (e.g. $\mathcal{M}'$ in Figure 1).

A PDFA $\mathcal{M}$ generates a stochastic language $\mathcal{D}_{\mathcal{M}}$. If it exists, the (unique) *path* for a word $w = a_0 \dots a_k$ belonging to Σ^* through a PDFA is a sequence $\langle (q_0, a_0), (q_1, a_1), \dots, (q_k, a_k) \rangle$, where $q_{i+1} = \delta(q_i, a_i)$. The probability a PDFA assigns to w is obtained by multiplying the transition probabilities with the final probability along w's path if it exists, and zero otherwise.

$$Pr_{\mathcal{D}_{\mathcal{M}}}(w) = \left(\prod_{i=0}^{k} T(q_i, a_i) \right) \cdot F(q_{k+1}) \quad (2)$$

$$\text{if } \hat{d}(q_0, w)\downarrow \text{ and 0 otherwise}$$

A stochastic language is *regular deterministic* iff there is a PDFA which generates it.

The *structural components* of a PDFA $\mathcal{M}$ is the deterministic finite-state automata (DFA) given by the states Q, alphabet Σ, transitions δ, and initial state q_0 of $\mathcal{M}$. By the *structure* of a PDFA, we mean its structural components.[1] Each PDFA $\mathcal{M}$ defines a family of distributions given by the possible instantiations of T and F satisfying Equation 1. These distributions have at most $|Q| \cdot (|\Sigma| + 1)$ parameters (since for each state there are $|\Sigma|$ possible transitions plus the possibility of finality.) These are, for all $q \in Q$ and $\sigma \in \Sigma$, the probabilities $T(q, \sigma)$ and $F(q)$. To make the connection to probability theory, we sometimes write these as $Pr(\sigma \mid q)$ and $Pr(\# \mid q)$, respectively.

We define the product of PDFAs in terms of *co-emission probabilities* (Vidal et al., 2005a). Let $\mathcal{M}_1 = \langle Q_1, \Sigma_1, q_{01}, \delta_1, F_1, T_1 \rangle$ and $\mathcal{M}_2 =$

[1] This is up to the renaming of states so PDFA with isomorphic structural components are said to have the same structure.

$\langle Q_2, \Sigma_2, q_{02}, \delta_2, F_2, T_2 \rangle$ be PDFAs. The probability that σ_1 is emitted from $q_1 \in Q_1$ at the same moment σ_2 is emitted from $q_2 \in Q_2$ is $CT(\sigma_1, \sigma_2, q_1, q_2) = T_1(q_1, \sigma_1) \cdot T_2(q_2, \sigma_2)$. Similarly, the probability that a word simultaneously ends at $q_1 \in Q_1$ and at $q_2 \in Q_2$ is $CF(q_1, q_2) = F_1(q_1) \cdot F_2(q_2)$.

Definition 1 *The* normalized co-emission product *of PDFAs $\mathcal{M}_1$ and $\mathcal{M}_2$ is $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 = \langle Q, \Sigma, q_0, \delta, F, T \rangle$ where*

1. *Q, q_0, and F are defined in terms of the standard DFA product over the state space $Q_1 \times Q_2$ (Hopcroft et al., 2001).*
2. *$\Sigma = \Sigma_1 \times \Sigma_2$*
3. *For all $\langle q_1, q_2 \rangle \in Q$ and $\langle \sigma_1, \sigma_2 \rangle \in \Sigma$, $\delta(\langle q_1, q_2 \rangle, \langle \sigma_1, \sigma_2 \rangle) = \langle q_1', q_2' \rangle$ iff $\delta_1(q_1, \sigma_1) = q_1'$ and $\delta_2(q_2, \sigma_2) = q_2'$.*[2]
4. *For all $\langle q_1, q_2 \rangle \in Q$,*
   (a) *let $Z(\langle q_1, q_2 \rangle) = CF(\langle q_1, q_2 \rangle) + \sum_{\langle \sigma_1, \sigma_2 \rangle \in \Sigma} CT(\sigma_1, \sigma_2, q_1, q_2)$ be the* normalization term*; and*
   (b) *$F(\langle q_1, q_2 \rangle) = \frac{CF(q_1, q_2)}{Z}$; and*
   (c) *for all $\langle \sigma_1, \sigma_2 \rangle \in \Sigma$, $T(\langle q_1, q_2 \rangle, \langle \sigma_1, \sigma_2 \rangle) = \frac{CT(\langle \sigma_1, \sigma_2, q_1, q_2 \rangle)}{Z}$*

In other words, the numerators of T and F are defined to be the co-emission probabilities, and division by Z ensures that $\mathcal{M}$ defines a well-formed probability distribution.[3] The normalized co-emission product effectively adopts a statistical independence assumption between the states of $\mathcal{M}_1$ and $\mathcal{M}_2$. If S is a list of PDFAs, we write $\bigotimes S$ for their product (note order of product is irrelevant up to renaming of the states).

The maximum likelihood (ML) estimation of regular deterministic distributions is a solved problem when the structure of the PDFA is known (Vidal et al., 2005a; Vidal et al., 2005b; de la Higuera, 2010). Let S be a finite sample of words drawn from a regular deterministic distribution $\mathcal{D}$. The problem is to estimate parameters T and F of $\mathcal{M}$ so that $\mathcal{D}_\mathcal{M}$ approaches $\mathcal{D}$ using the widely-adopted ML criterion (Equation 3).

$$(\hat{T}, \hat{F}) = \underset{T,F}{\operatorname{argmax}} \left(\prod_{w \in S} Pr_\mathcal{M}(w) \right) \quad (3)$$

It is well-known that if $\mathcal{D}$ is generated by some PDFA $\mathcal{M}'$ with the same structural components as $\mathcal{M}$, then the *ML estimate of S with respect to $\mathcal{M}$* guarantees that $\mathcal{D}_\mathcal{M}$ approaches $\mathcal{D}$ as the size of S goes to infinity (Vidal et al., 2005a; Vidal et al., 2005b; de la Higuera, 2010).

Finding the ML estimate of a finite sample S with respect to $\mathcal{M}$ is simple provided $\mathcal{M}$ is deterministic with known structural components. Informally, the corpus is passed through the PDFA, and the paths of each word through the corpus are tracked to obtain counts, which are then normalized by state. Let $\mathcal{M} = \langle Q, \Sigma, \delta, q_0, F, T \rangle$ be the PDFA whose parameters F and T are to be estimated. For all states $q \in Q$ and symbols $\sigma \in \Sigma$, The ML estimation of the probability of $T(q, \sigma)$ is obtained by dividing the number of times this transition is used in parsing the sample S by the number of times state q is encountered in the parsing of S. Similarly, the ML estimation of $F(q)$ is obtained by calculating the relative frequency of state q being final with state q being encountered in the parsing of S. For both cases, the division is *normalizing*; i.e. it guarantees that there is a well-formed probability distribution at each state. Figure 1 illustrates the counts obtained for a machine $\mathcal{M}$ with sample $S = \{abca\}$.[4] Figure 1 shows a DFA with counts and the PDFA obtained after normalizing these counts.

3 Strictly local distributions

In formal language theory, *strictly k-local* languages occupy the bottom rung of a subregular hierarchy which makes distinctions on the basis of contiguous subsequences (McNaughton and Papert, 1971; Rogers and Pullum, to appear; Rogers et al., 2009). They are also the categorical counterpart to stochastic languages describable with n-gram models (where $n = k$) (Garcia et al., 1990; Jurafsky and Martin, 2008). Since stochastic languages are distributions, we refer to strictly k-local stochastic languages as strictly k-local distri-

[2] Note that restricting δ to cases when $\sigma_1 = \sigma_2$ obtains the standard definition of $\delta = \delta_1 \times \delta_2$ (Hopcroft et al., 2001). The reason we maintain two alphabets becomes clear in §4.

[3] $Z(\langle q_1, q_2 \rangle)$ is less than one whenever either $F_1(q_1)$ or $F_2(q_2)$ are neither zero nor one.

[4] Technically, $\mathcal{M}$ is neither a simple DFA or PDFA; rather, it has been called a Frequency DFA. We do not formally define them here, see de la Higuera (2010).

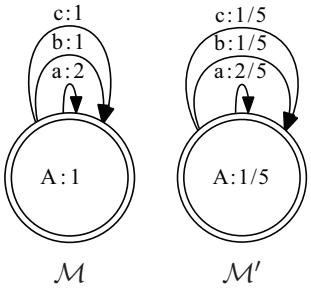

Figure 1: $\mathcal{M}$ shows the counts obtained by parsing it with sample $S = \{abca\}$. $\mathcal{M}'$ shows the probabilities obtained after normalizing those counts.

butions (SLD$_k$). We illustrate with SLD$_2$ (bigram models) for ease of exposition.

For an alphabet Σ, SL$_2$ distributions have $(|\Sigma| + 1)^2$ parameters. These are, for all $\sigma, \tau \in \Sigma \cup \{\#\}$, the probabilities $Pr(\sigma \mid \tau)$. The probability of $w = \sigma_1 \ldots \sigma_n$ is given in Equation 4.

$$Pr(w) \stackrel{\text{def}}{=} Pr(\sigma_1 \mid \#) \times Pr(\sigma_2 \mid \sigma_1) \times \ldots \times Pr(\# \mid \sigma_n) \quad (4)$$

PDFA representations of SL$_2$ distributions have the following structure: $Q = \Sigma \cup \{\#\}$, $q_0 = \#$, and for all $q \in Q$ and $\sigma \in \Sigma$, it is the case that $\delta(q, \sigma) = \sigma$.

As an example, the DFA in Figure 2 provides the structure of PDFAs which recognize SL$_2$ distributions with $\Sigma = \{a, b, c\}$. Plainly, the parameters of the model are given by assigning probabilities to each transition and to the ending at each state. In fact, for all $\sigma \in \Sigma$ and $\tau \in \Sigma \cup \{\#\}$, $Pr(\sigma \mid \tau)$ is $T(\tau, \sigma)$ and $Pr(\# \mid \tau)$ is $F(\tau)$. It follows that the probability of a particular path through the model corresponds to Equation 4. The structure of a SL$_2$ distribution for alphabet Σ is given by $\mathcal{M}_{\text{SL2}}(\Sigma)$.

Additionally, given a finite sample $S \subset \Sigma^*$, the ML estimate of S with respect to the family of distributions describable with $\mathcal{M}_{\text{SL2}}(\Sigma)$ is given by counting the parse of S through $\mathcal{M}_{\text{SL2}}(\Sigma)$ and then normalizing as described in §2. This is equivalent to the procedure described in Jurafsky and Martin (2008, chap. 4).

4 Feature-based distributions

This section first introduces feature systems. Then it defines feature-based SL$_2$ distributions which make the strong independence assumption that no two features interact. It explains how to find the ML estimate of samples with respect to such distributions. This section closes by identifying kinds of featural interactions in phonotactic patterns, and discusses how such interactions can be addressed within this framework.

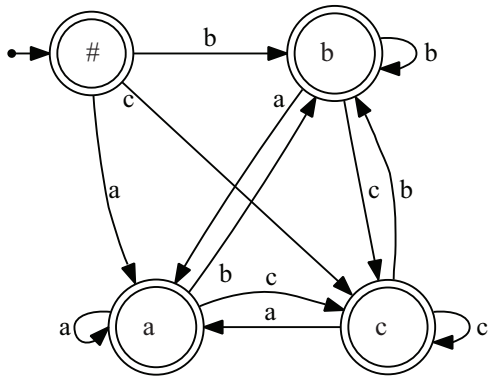

Figure 2: $\mathcal{M}_{\text{SL2}}(\{a, b, c\})$ represents the structure of SL$_2$ distributions when $\Sigma = \{a, b, c\}$.

	F	G
a	+	-
b	+	+
c	-	+

Table 1: An example of a feature system with $\Sigma = \{a, b, c\}$ and two features F and G.

4.1 Feature systems

Assume the elements of the alphabet share properties, called features. For concreteness, let each feature be a total function $F : \Sigma \to \mathbb{V}_F$, where the codomain $\mathbb{V}_F$ is a finite set of *values*. A finite vector of features $\mathbb{F} = \langle F_1, \ldots, F_n \rangle$ is called a *feature system*. Table 1 provides an example of a feature system with $\mathbb{F} = \langle F, G \rangle$ and values $\mathbb{V}_F = \mathbb{V}_G = \{+, -\}$.

We extend the domain of all features $F \in \mathbb{F}$ to Σ^+, so that $F(\sigma_1 \ldots \sigma_n) = F(\sigma_1) \ldots F(\sigma_n)$. For example, using the feature system in Table 1, $F(abc) = + + -$ and $G(abc) = - + +$. We also extend the domain of F to all languages: $F(L) = \cup_{w \in L} f(w)$. We also extend the notation so that $\mathbb{F}(\sigma) = \langle F_1(\sigma), \ldots, F_n(\sigma) \rangle$. For example, $\mathbb{F}(c) = \langle -F, +G \rangle$ (feature indices are included for readability).

For feature $F : \Sigma \to \mathbb{V}_F$, let F^{-1} be the inverse function with domain V_F and codomain $\mathcal{P}(\Sigma)$. For example in Table 1, $G^{-1}(+) = \{b, c\}$. $\mathbb{F}^{-1}$ is similarly defined, i.e. $\mathbb{F}^{-1}(\langle -F, +G \rangle) = \{c\}$.

If, for all arguments $\vec{v}$, $\mathbb{F}^{-1}(\vec{v})$ is nonempty then the feature system is *exhaustive*. If, for all arguments $\vec{v}$ such that $\mathbb{F}^{-1}(\vec{v})$ is nonempty, it is the case that $|\mathbb{F}^{-1}(\vec{v})| = 1$ then the feature system is *distinctive*. E.g. the feature system in Table 1 in not exhaustive since $\mathbb{F}^{-1}(\langle -F, -G\rangle) = \emptyset$, but it is distinctive since where $\mathbb{F}^{-1}$ is nonempty, it picks out exactly one element of the alphabet.

Generally, phonological feature systems for a particular language are distinctive but not exhaustive. Any feature system $\mathbb{F}$ can be made exhaustive by adding finitely many symbols to the alphabet (since $\mathbb{F}$ is finite). Let Σ' denote an alphabet obtained by adding to Σ the fewest symbols which make $\mathbb{F}$ exhaustive.

Each feature system also defines a set of indicator functions $\mathbb{VF} = \bigcup_{f\in\mathbb{F}}(\mathbb{V}_f \times \{f\})$ with domain Σ such that $\langle v, f\rangle(\sigma) = 1$ iff $f(\sigma) = v$ and 0 otherwise. In the example in Table 1, $\mathbb{VF} = \{+F, -F, +G, -G\}$ (omitting angle braces for readability). For all $f \in \mathbb{F}$, the set $\mathbb{VF}_f$ is the $\mathbb{VF}$ restricted to f. So continuing our example, $\mathbb{VF}_F = \{+F, -F\}$.

4.2 Feature-based distributions

We now define feature-based SL_2 distributions under the strong independence assumption that no two features interact. For feature system $\mathbb{F} = \langle F_1 \dots F_n\rangle$, there are n PDFAs, one for each feature. The normalized co-emission product of these PDFAs essentially defines the distribution. For each F_i, the structure of its PDFA is given by $\mathcal{M}_{\text{SL2}}(\mathbb{V}_{F_i})$. For example, $\mathcal{M}_F = \mathcal{M}_{\text{SL2}}(\mathbb{V}_F)$ and $\mathcal{M}_G = \mathcal{M}_{\text{SL2}}(\mathbb{V}_G)$ in figures 3 and 4 illustrate the finite-state representation of feature-based SL_2 distributions given the feature system in Table 1.[5] The states of each machine make distinctions according to features F and G, respectively. The parameters of these distributions are given by assigning probabilities to each transition and to the ending at each state (except for $Pr(\# \mid \#)$).[6]

Thus there are $2|\mathbb{VF}| + \sum_{F\in\mathbb{F}}|\mathbb{VF}_F|^2 + 1$ parameters for feature-based SL_2 distributions. For example, the feature system in Table 1 defines a distribution with $2\cdot 4 + 2^2 + 2^2 + 1 = 17$ parameters, which include $Pr(\# \mid +F)$, $Pr(+F \mid \#)$, $Pr(+F \mid +F)$, $Pr(+F \mid -F)$, …, the G equivalents, and $Pr(\# \mid \#)$. Let $\text{SLD2}_{\mathbb{F}}$ be the family of distributions given by all possible parameter settings (i.e. all possible probability assignments for each $\mathcal{M}_{\text{SL2}}(\mathbb{V}_{F_i})$ in accordance with Equation 1.)

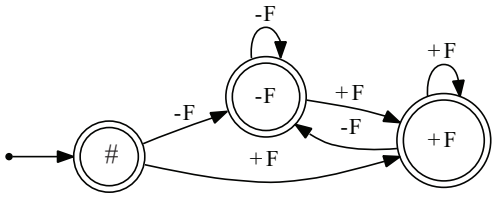

Figure 3: $\mathcal{M}_F$ represents a SL_2 distribution with respect to feature F.

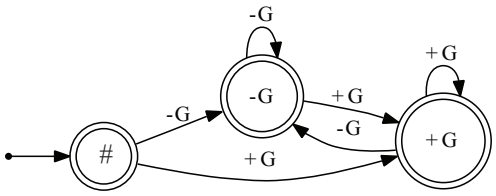

Figure 4: $\mathcal{M}_G$ represents a SL_2 distribution with respect to feature G.

The normalized co-emission product defines the feature-based distribution. For example, the structure of the product of $\mathcal{M}_F$ and $\mathcal{M}_G$ is shown in Figure 5.

As defined, the normalized co-emission product can result in states and transitions that cannot be interpreted by non-exhaustive feature systems. An example of this is in Figure 5 since $\langle -F, -G\rangle$ is not interpretable by the feature system in Table 1. We make the system exhaustive by letting $\Sigma' = \Sigma \cup \{d\}$ and setting $\mathbb{F}(d) = \langle -F, -G\rangle$.

What is the probability of a given b in the feature-based model? According to the normalized co-emission product (Defintion 1), it is

$$Pr(a \mid b) = Pr(\langle +F, -G\rangle \mid \langle +F, +G\rangle) = \frac{Pr(+F \mid +F)\cdot Pr(-G \mid +G)}{Z}$$

where $Z = Z(\langle +F, +G\rangle)$ equals

$$\sum_{\sigma\in\Sigma'} \Pr(F(\sigma) \mid +F)\cdot \Pr(G(\sigma) \mid +G) + (\Pr(\# \mid +F)\cdot \Pr(\# \mid +G)$$

Generally, for an *exhuastive* distinctive feature system $\mathbb{F} = \langle F_1, \dots, F_n\rangle$, and for all $\sigma, \tau \in \Sigma$,

[5] For readability, featural information in the states and transitions is included in these figures. By definition, the states and transitions are only labeled with elements of $\mathbb{V}_F$ and $\mathbb{V}_G$, respectively. In this case, that makes the structures of the two machines identical.

[6] It is possible to replace $Pr(\# \mid \#)$ with two parameters, $Pr(\# \mid \#_F)$ $Pr(\# \mid \#_G)$, but for ease of exposition we do not pursue this further.

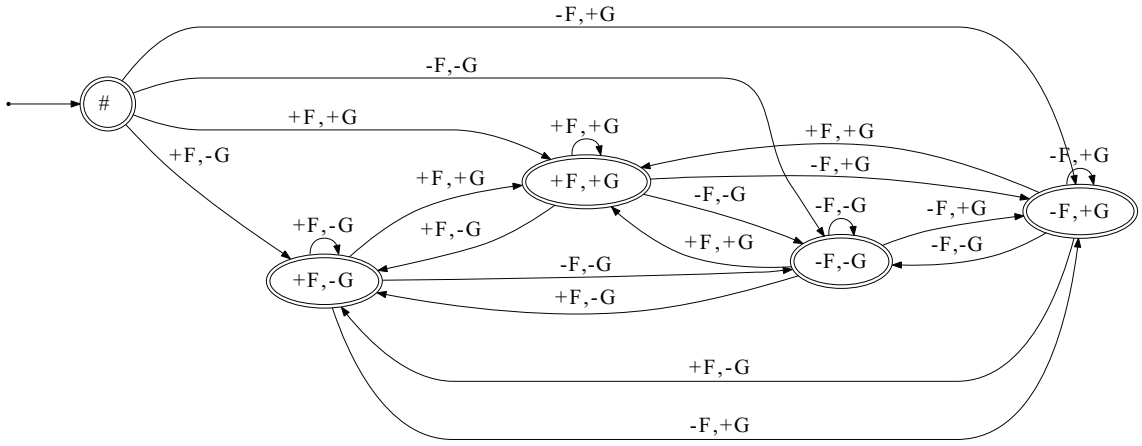

Figure 5: The structure of the product of $\mathcal{M}_F$ and $\mathcal{M}_G$.

the $Pr(\sigma \mid \tau)$ is given by Equation 5. First, the normalization term is provided. Let

$$Z(\tau) = \sum_{\sigma \in \Sigma} \left(\prod_{1 \leq i \leq n} Pr(F_i(\sigma) \mid F_i(\tau)) \right) + \prod_{1 \leq i \leq n} Pr(\# \mid F_i(\tau))$$

Then

$$Pr(\sigma \mid \tau) = \frac{\prod_{1 \leq i \leq n} Pr(F_i(\sigma) \mid F_i(\tau))}{Z(\tau)} \quad (5)$$

The probabilities $Pr(\sigma \mid \#)$ and $Pr(\# \mid \tau)$ are similarly decomposed into featural parameters. Finally, like SL_2 distributions, the probability of a word $w \in \Sigma^*$ is given by Equation 4. We have thus proved the following.

Theorem 1 *The parameters of a feature-based* SL_2 *distribution define a well-formed probability distribution over* Σ^*.

Proof It is sufficient to show for all $\tau \in \Sigma \cup \{\#\}$ that $\sum_{\sigma \in \Sigma \cup \{\#\}} Pr(\sigma \mid \tau) = 1$ since in this case, Equation 4 yields a well-formed probability distribution over Σ^*. This follows directly from the definition of the normalized co-emission product (Definition 1). □

The normalized co-emission product adopts a statistical independence assumption, which here is between features since each machine represents a single feature. For example, consider $Pr(a \mid b) = Pr(\langle -F, +G \rangle \mid \langle +F, +G \rangle)$. The probability $Pr(\langle -F, +G \rangle \mid \langle +F, +G \rangle)$ cannot be arbitrarily different from the probabilities $Pr(-F \mid +F)$ and $Pr(+G \mid +G)$; it is not an independent parameter. In fact, because $Pr(a \mid b)$ is computed directly as the normalized product of parameters $Pr(-F \mid +F)$ and $Pr(+G \mid +G)$, the assumption is that the features F and G do not interact. In other words, this model describes exactly the state of affairs one expects if there is no statistical interaction between phonological features. In terms of inference, this means if one sound is observed to occur in some context (at least contexts distinguishable by SL_2 models), then similar sounds (i.e. those that share many of its featural values) are expected to occur in this context as well.

4.3 ML estimation

The ML estimate of feature-based SL_2 distributions is obtained by counting the parse of a sample through each feature machine, and normalizing the results. This is because the parameters of the distribution are the probabilities on the feature machines, whose product determines the actual distribution. The following theorem follows immediately from the PDFA representation of feature-based SL_2 distributions.

Theorem 2 *Let* $\mathbb{F} = \langle F_1, \ldots F_n \rangle$ *and let* $\mathcal{D}$ *be described by* $\mathcal{M} = \bigotimes_{1 \leq i \leq n} \mathcal{M}_{\mathrm{SL2}}(\mathbb{V}_{Fi})$. *Consider a finite sample* S *drawn from* $\mathcal{D}$. *Then the ML estimate of* S *with respect to* $\mathrm{SLD2}_{\mathbb{F}}$ *is obtained by finding, for each* $F_i \in \mathbb{F}$, *the ML estimate of* $F_i(S)$ *with respect to* $\mathcal{M}_{\mathrm{SL2}}(\mathbb{V}_{Fi})$.

Proof The ML estimate of S with respect to $\mathrm{SLD2}_{\mathbb{F}}$ returns the parameter values that maximize the likelihood of S within the family $\mathrm{SLD2}_{\mathbb{F}}$. The parameters of $\mathcal{D} \in \mathrm{SLD2}_{\mathbb{F}}$ are found on the

states of each $\mathcal{M}_{\text{SL2}}(\mathbb{V}_{F_i})$. By definition, each $\mathcal{M}_{\text{SL2}}(\mathbb{V}_{F_i})$ describes a probability distribution over $F_i(\Sigma^*)$, as well as a family of distributions. Therefore finding the MLE of S with respect to SLD2$_{\mathbb{F}}$ means finding the MLE estimate of $F_i(S)$ with respect to each $\mathcal{M}_{\text{SL2}}(\mathbb{V}_{F_i})$.

Optimizing the ML estimate of $F_i(S)$ for each $\mathcal{M}_i = \mathcal{M}_{\text{SL2}}(\mathbb{V}_{F_i})$ means that as $|F_i(S)|$ increases, the estimates $\hat{T}_{\mathcal{M}_i}$ and $\hat{F}_{\mathcal{M}_i}$ approach the true values $T_{\mathcal{M}_i}$ and $F_{\mathcal{M}_i}$. It follows that as $|S|$ increases, $\hat{T}_{\mathcal{M}}$ and $\hat{F}_{\mathcal{M}}$ approach the true values of $T_{\mathcal{M}}$ and $F_{\mathcal{M}}$ and consequently $\mathcal{D}_{\mathcal{M}}$ approaches $\mathcal{D}$. □

4.4 Discussion

Feature-based models can have significantly fewer parameters than segment-based models. Consider binary feature systems, where $|\mathbb{VF}| = 2|\mathbb{F}|$. An exhaustive feature system with 10 binary features describes an alphabet with 1024 symbols. Segment-based bigram models have $(1024{+}1)^2 = 1,050,625$ parameters, but the feature-based one only has $40 + 40 + 1 = 81$ parameters! Consequently, much less training data is required to accurately estimate the parameters of the model.

Another way of describing this is in terms of expressivity. For given feature system, feature-based SL$_2$ distributions are a proper subset of SL$_2$ distributions since, as the the PDFA representations make clear, every feature-based distribution can be described by a segmental bigram model, but not vice versa. The fact that feature-based distributions have potentially far fewer parameters is a reflection of the restrictive nature of the model. The statistical independence assumption constrains the system in predictable ways. The next section shows exactly what feature-based generalization looks like under these assumptions.

5 Examples

This section demonstrates feature-based generalization by comparing it with segment-based generalization, using a small corpus $S = \{aaab, caca, acab, cbb\}$ and the feature system in Table 1. Tables 2 and 3 show the results of ML estimation of S with respect to segment-based SL$_2$ distributions (unsmoothed bigram model) and feature-based SL$_2$ distributions, respectively. Each table shows the $Pr(\sigma \mid \tau)$ for all $\sigma, \tau \in \{a, b, c, d, \#\}$ (where $\mathbb{F}(d) = \langle -F, -G \rangle$), for ease of comparison.

P($\sigma \mid \tau$)		σ				
		a	b	c	d	#
τ	a	0.29	0.29	0.29	0.	0.14
	b	0.	0.25	0.	0.	0.75
	c	0.75	0.25	0.	0.	0.
	d	0.	0.	0.	0.	0.
	#	0.5	0.	0.5	0.	0.

Table 2: ML estimates of parameters of segment-based SL$_2$ distributions.

P($\sigma \mid \tau$)		σ				
		a	b	c	d	#
τ	a	0.22	0.43	0.17	0.09	0.09
	b	0.32	0.21	0.09	0.13	0.26
	c	0.60	0.40	0.	0	0.
	d	0.33	0.67	0	0	0
	#	0.25	0.25	0.25	0.25	0.

Table 3: ML estimates of parameters of feature-based SL$_2$ distributions.

Observe the sharp divergence between the two models in certain cells. For example, no words begin with b in the sample. Hence the segment-based ML estimates of $Pr(b \mid \#)$ is zero. Conversely, the feature-based ML estimate is nonzero because b, like a, is +F, and b, like c, is +G, and both a and c begin words. Also, notice nonzero probabilities are assigned to d occuring after a and b. This is because $\mathbb{F}(d) = \langle -F, -G \rangle$ and the following sequences all occur in the corpus: [+F][-F] (ac), [+G][-G] (ca), and [-G][-G] (aa). On the other hand, zero probabilities are assigned to d ocurring after c and d because there are no cc sequences in the corpus and hence the probability of [-F] occuring after [-F] is zero.

This simple example demonstrates exactly how the model works. Generalizations are made on the basis of individual features, not individual symbols. In fact, segments are truly epiphenomenal in this model, as demonstrated by the nonzero probabilties assigned to segments outside the original feature system (here, this is d). To sum up, this model captures exactly the idea that the distribution of segments is conditioned on the distributions of its features.

6 Featural interaction

In many empirical cases of interest, features do interact, which suggests the strong independence assumption is incorrect for modeling phonotactic learning.

There are at least four kinds of featural interaction. First, different features may be prohibited from occuring simultaneously in certain contexts. As an example of the first type consider the fact that both velars and nasal sounds occur word-initially in English, but the velar nasal may not. Second, specific languages may prohibit different features from simultaneously occuring in all contexts. In English, for example, there are syllabic sounds and obstruents but no syllabic obstruents. Third, different features may be universally incompatible: e.g. no vowels are both [+high] and [+low]. The last type of interaction is that different features may be prohibited from occuring syntagmatically. For example, some languages prohibit voiceless sounds from occuring after nasals.

Although the independence assumption is too strong, it is still useful. First, it allows researchers to quantify the extent to which data can be explained without invoking featural interaction. For example, following Hayes and Wilson (2008), we may be interested in how well human acceptability judgements collected by Scholes (1966) can be explained if different features do not interact. After training the feature-based SL_2 model on a corpus of word initial onsets adapted from the CMU pronouncing dictionary (Hayes and Wilson, 2008, 395-396) and using a standard phonological feature system (Hayes, 2009, chap. 4), it achieves a correlation (Spearman's r) of 0.751.[7] In other words, roughly three quarters of the acceptability judgements are explained without relying on featural interaction (or segments).

Secondly, the incorrect predictions of the model are in principle detectable. For example, recall that English has word-inital velars and nasals, but no word-inital velar nasals. A one-cell chi-squared test can determine whether the observed number of [#ŋ] is significantly below the expected number according to the feature-based distribution, which could lead to a new parameter being adopted to describe the interaction of the [dorsal] and [nasal] features word-initially. The details of these procedures are left for future research and are likely to draw from the rich literature on Bayesian networks (Pearl, 1989; Ghahramani, 1998).

More important, however, is this framework allows researchers to construct the independence assumptions they want into the model in at least two ways. First, universally incompatible features can be excluded. For example, suppose [-F] and [-G] in the feature system in Table 1 are anatomically incompatible like [+low] and [+high]. If desired, they can be excluded from the model essentially by zeroing out any probability mass assigned to such combinations and re-normalizing.

Second, models can be defined where multiple features are permitted to interact. For example, suppose features F and G from Table 1 are embedded in a larger feature system. The machine in Figure 5 can be defined to be a *factor* of the model, and now interactions between F and G will be learned, including syntagmatic ones. The flexibility of the framework and the generality of the normalized co-emission product allow researchers to consider feature-based distributions which allow any two features to interact but which prohibit three-feature interactions, or which allow any three features to interact but which prohibit four-feature interactions, or models where only certain features are permitted to interact but not others (perhaps because they belong to the same node in a feature geometry (Clements, 1985; Clements and Hume, 1995).[8]

7 Hayes and Wilson (2008)

This section introduces the Hayes and Wilson (2008) (henceforth HW) phonotactic learner and shows that the contribution features play in generalization is not as clear as previously thought.

HW propose an inductive model which acquires a maxent grammar defined by weighted constraints. Each constraint is described as a sequence of natural classes using phonological features. The constraint format also allows reference to word boundaries and at most one complement class. (The complement class of $S \subseteq \Sigma$ is Σ/S.) For example, the constraint

*#[^ -voice,+anterior,+strident][-approximant]

means that in word-initial C_1C_2 clusters, if C_2 is a nasal or obstruent, then C_1 must be [s].

[7] We use the feature chart in Hayes (2009) because it contains over 150 IPA symbols (and not just English phonemes). Featural combinations not in the chart were assumed to be impossible (e.g. [+high,+low]) and were zeroed out.

[8] Note if all features are permitted to interact, this yields the segmental bigram model.

Hayes and Wilson maxent models	r
features & complement classes	0.946
no features & complement classes	0.937
features & no complement classes	0.914
no features & no complement classes	0.885

Table 4: Correlations of different settings versions of HW maxent model with Scholes data.

HW report that the model obtains a correlation (Spearman's r) of 0.946 with blick test data from Scholes (1966). HW and Albright (2009) attribute this high correlation to the model's use of natural classes and phonological features. HW also report that when the model is run without features, the grammar obtained scores an r value of only 0.885, implying that the gain in correlation is due specifically to the use of phonological features.

However, there are two relevant issues. The first is the use of complement classes. If features are not used but complement classes are (in effect only allowing the model to refer to single segments and the complements of single segments, e.g. [t] and [ˆt]) then in fact the grammar obtained scores an r value of 0.936, a result comparable to the one reported.[9] Table 4 shows the r values obtained by the HW learner under different conditions. Note we replicate the main result of $r = 0.946$ when using both features and complement classes.[10]

This exercise reveals that phonological features play a smaller role in the HW phonotactic learner than previously thought. Features are helpful, but not as much as complement classes of single segments (though features with complement classes yields the best result by this measure).

The second issue relates to the first: the question of whether additional parameters are worth the gain in empirical coverage. Wilson and Obdeyn (2009) provide an excellent discussion of the model comparison literature and provide a rigorous comparative analysis of computational modeleling of OCP restrictions. Here we only raise the questions and leave the answers to future research. Compare the HW learners in the first two rows in Table 4. Is the ~ 0.01 gain in r score worth the additional parameters which refer to phonologically natural classes? Also, the feature-based SL_2 model in §4 only receives an r score of 0.751, much lower than the results in Table 4. Yet this model has far fewer parameters not only because the maxent models in Table 4 keep track of trigrams, but also because of its strong independence assumption. As mentioned, this result is informative because it reveals how much can be explained without featural interaction. In the context of model comparison, this particular model provides an inductive baseline against which the utility of additional parameters invoking featural interaction ought to be measured.

[9] Examination of the output grammar reveals heavy reliance on the complement class [ˆs], which is not surprising given the discussion of [sC] clusters in HW.

[10] This software is available on Bruce Hayes' webpage: `http://www.linguistics.ucla.edu/people/hayes/Phonotactics/index.htm`.

8 Conclusion

The current proposal explicitly embeds the Jakobsonian hypothesis that the primitive unit of phonology is the phonological feature into a phonotactic learning model. While this paper specifically shows how to integrate features into n-gram models to describe feature-based strictly n-local distributions, these techniques can be applied to other regular deterministic distributions, such as strictly k-piecewise models, which describe long-distance dependencies, like the ones found in consonant and vowel harmony (Heinz, to appear; Heinz and Rogers, 2010).

In contrast to models which assume that all features potentially interact, a baseline model was specifically introduced under the assumption that no two features interact. In this way, the "bottom-up" approach to feature-based generalization shifts the focus of inquiry to the featural interactions necessary (and ultimately sufficient) to describe and learn phonotactic patterns. The framework introduced here shows how researchers can study feature interaction in phonotactic models in a systematic, transparent way.

Acknowledgments

We thank Bill Idsardi, Tim O'Neill, Jim Rogers, Robert Wilder, Colin Wilson and the U. of Delaware's phonology/phonetics group for valuable discussion. Special thanks to Mark Ellison for helpful comments, to Adam Albright for illuminating remarks on the types of featural interaction in phonotactic patterns, and to Jason Eisner for bringing to our attention FHMMs and other related work.

References

Adam Albright. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1):9–41.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.

G.N. Clements and Elizabeth V. Hume. 1995. The internal organization of speech sounds. In John A. Goldsmith, editor, *The handbook of phonological theory*, chapter 7. Blackwell, Cambridge, MA.

George N. Clements. 1985. The geometry of phonological features. *Phonology Yearbook*, 2:225–252.

Colin de la Higuera. 2010. *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press.

Markus Dreyer and Jason Eisner. 2009. Graphical models over multiple strings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 101–110, Singapore, August.

Markus Dreyer, Jason R. Smith, and Jason Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1080–1089, Honolulu, October.

Pedro Garcia, Enrique Vidal, and José Oncina. 1990. Learning locally testable languages in the strict sense. In *Proceedings of the Workshop on Algorithmic Learning Theory*, pages 325–338.

Zoubin Ghahramani and Michael I. Jordan. 1997. Factorial hidden markov models. *Machine Learning*, 29(2):245–273.

Zoubin Ghahramani. 1998. Learning dynamic bayesian networks. In *Adaptive Processing of Sequences and Data Structures*, pages 168–197. Springer-Verlag.

Daniel Gildea and Daniel Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics*, 24(4).

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Bruce Hayes. 2009. *Introductory Phonology*. Wiley-Blackwell.

Jeffrey Heinz and James Rogers. 2010. Estimating strictly piecewise distributions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Jeffrey Heinz. to appear. Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4).

John Hopcroft, Rajeev Motwani, and Jeffrey Ullman. 2001. *Introduction to Automata Theory, Languages, and Computation*. Boston, MA: Addison-Wesley.

Roman Jakobson, C. Gunnar, M. Fant, and Morris Halle. 1952. *Preliminaries to Speech Analysis*. MIT Press.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, Upper Saddle River, NJ, 2nd edition.

Robert McNaughton and Seymour Papert. 1971. *Counter-Free Automata*. MIT Press.

Elliot Moreton. 2008. Analytic bias and phonological typology. *Phonology*, 25(1):83–127.

Judea Pearl. 1989. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman.

James Rogers and Geoffrey Pullum. to appear. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information*.

James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. 2009. On languages piecewise testable in the strict sense. In *Proceedings of the 11th Meeting of the Assocation for Mathematics of Language*.

Lawrence K. Saul and Michael I. Jordan. 1999. Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, 37(1):75–87.

Robert J. Scholes. 1966. *Phonotactic grammaticality*. Mouton, The Hague.

Enrique Vidal, Franck Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. 2005a. Probabilistic finite-state machines-part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025.

Enrique Vidal, Frank Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. 2005b. Probabilistic finite-state machines-part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1026–1039.

Colin Wilson and Marieke Obdeyn. 2009. Simplifying subsidiary theory: statistical evidence from arabic, muna, shona, and wargamay. Johns Hopkins University.

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5):945–982.