Learning gradient long-distance phonotactics by estimating strictly piecewise distributions

Jeffrey Heinz
heinz@udel.edu

University of Delaware

University of Alberta
February 13, 2010

Collaborators

James Rogers
Earlham College

Overview

Gradient phonotactic long-distance dependencies are provably feasibly learnable

1. without tiers
2. without a concept of similarity
3. without the additional structure provided by OT or P&P frameworks
4. with a naturally structured hypothesis space
5. with a learner whose behavior is provably correct

Overview

Gradient phonotactic long-distance dependencies are provably feasibly learnable

1. without tiers
2. without a concept of similarity
3. without the additional structure provided by OT or P&P frameworks
4. with a naturally structured hypothesis space
5. with a learner whose behavior is provably correct

Goals of this talk

1. Show how gradient constraints like $*a \ldots b$ can be learned from positive evidence
2. Demonstrate the learner on various corpora
3. Address issues concerning typology, similarity, and tiers
   3.1 Since the learner can learn long-distance dependencies between unlike segments, aren't SP distributions too unrestrictive?
   3.2 Since the learner can't learn adjacency dependencies, aren't SP distributions too restrictive?
   3.3 What about phonological features?
   3.4 What about phonological tiers?

Goals of this talk

1. Show how gradient constraints like $*a \ldots b$ can be learned from positive evidence
2. Demonstrate the learner on various corpora
3. Address issues concerning typology, similarity, and tiers
    3.1 Since the learner can learn long-distance dependencies between unlike segments, aren't SP distributions too unrestrictive?
    3.2 Since the learner can't learn adjacency dependencies, aren't SP distributions too restrictive?
    3.3 What about phonological features?
    3.4 What about phonological tiers?

Goals of this talk

1. Show how gradient constraints like $*a \ldots b$ can be learned from positive evidence
2. Demonstrate the learner on various corpora
3. Address issues concerning typology, similarity, and tiers
   3.1 Since the learner can learn long-distance dependencies between unlike segments, aren't SP distributions too unrestrictive?
   3.2 Since the learner can't learn adjacency dependencies, aren't SP distributions too restrictive?
   3.3 What about phonological features?
   3.4 What about phonological tiers?

Goals of this talk

1. Show how gradient constraints like $*a \dots b$ can be learned from positive evidence
2. Demonstrate the learner on various corpora
3. Address issues concerning typology, similarity, and tiers
   3.1 Since the learner can learn long-distance dependencies between unlike segments, aren't SP distributions too unrestrictive?
   3.2 Since the learner can't learn adjacency dependencies, aren't SP distributions too restrictive?
   3.3 What about phonological features?
   3.4 What about phonological tiers?

Goals of this talk

1. Show how gradient constraints like $*a \ldots b$ can be learned from positive evidence
2. Demonstrate the learner on various corpora
3. Address issues concerning typology, similarity, and tiers
   3.1 Since the learner can learn long-distance dependencies between unlike segments, aren't SP distributions too unrestrictive?
   3.2 Since the learner can't learn adjacency dependencies, aren't SP distributions too restrictive?
   3.3 What about phonological features?
   3.4 What about phonological tiers?

Outline

Outline

Long-distance dependencies

SP Distributions

Estimating SP Distributions

Demo

Issues

Long-distance dependencies in phonology

1. **Consonantal harmony**

   (Jensen 1974, Odden 1994, Hansson 2001, Rose and Walker 2004, and many others)

2. **Vowel harmony**

   (Ringen 1988, Baković 2000, and many others)

Sibilant Harmony example from Samala (Ineseño Chumash)

[ʃtojonowonowaʃ] 'it stood upright' (Applegate 1972:72)

cf. *[**s**tojonowonowa**ʃ**] and

cf. *[**ʃ**tojonowonowa**s**]

Sibilant Harmony example from Samala (Ineseño Chumash)

[ʃtojonowonowaʃ] ‘it stood upright’ (Applegate 1972:72)

cf. *[**s**tojonowonowa**ʃ**] and

cf. *[**ʃ**tojonowonowa**s**]

Hypothesis:

*[stojonowonowaʃ] and *[ʃtojonowonowas] are ill-formed because the *dis*contiguous subsequences *sʃ* and *ʃs* are ill-formed.

(Heinz 2007, Rogers et. al 2009, Heinz to appear)

Outline

Main Idea

$$\boxed{w = a_1 a_2 \dots a_n}$$

Markov Assumption
(e.g. bigram model = strictly 2-local distribution)

$$\begin{aligned} Pr(w) \stackrel{\text{def}}{=}\ & Pr(a_1 \mid \#) \ \times \ Pr(a_2 \mid a_1 \cdot\,) \ \times \dots \\ & \times Pr(a_n \mid a_{n-1} \cdot\,) \times \ Pr(\# \mid a_n) \end{aligned}$$

Our Assumption (strictly 2-piecewise distribution)

$$\begin{aligned} Pr(w) \stackrel{\text{def}}{=}\ & Pr(a_1 \mid \#) \times \ Pr(a_2 \mid a_1 <) \dots \\ & \times Pr(a_n \mid a_1, \dots, a_{n-1} <) \\ & \times Pr(\# \mid a_1, \dots a_n <) \end{aligned} \tag{1}$$

Strictly 2-Piecewise distributions

w=ʃtojonowonowaʃ

Our Assumption (strictly 2-piecewise model)

$$Pr(w) \stackrel{\text{def}}{=} Pr(a_1 \mid \#) \times Pr(a_2 \mid a_1 <) \ldots \\ \times Pr(a_n \mid a_1, \ldots, a_{n-1} <) \\ \times Pr(\# \mid a_1, \ldots a_n <) \quad (1)$$

Key Ideas:

1. Pr(ʃ | ʃ,t,o,y,w,n,a $<$) $\gg$ Pr(s | ʃ,t,o,y,w,n,a $<$)
2. Pr(s | ʃ,t,o,y,w,n,a $<$) is impossibly low because Pr(s | ʃ$<$) is impossibly low.

Problem and Solution

What is $Pr(a \mid S)$?
There are $2^{|\Sigma|}$ distinct sets S which suggests there are too many(!) independent parameters in the model.

Solution: $Pr(a \mid S <)$ is a function of $Pr(a \mid s <)$ for all $s \in S$

$$Pr(a \mid S <) \stackrel{\text{def}}{=} \frac{\prod_{s \in S} Pr(a|s<)}{\sum_{a' \in \Sigma \cup \{\#\}} \prod_{s \in S} Pr(a'|s)} \qquad (2)$$

Theorem (Heinz and Rogers, in prep)

Equations (1) and (2) guarantee a well-formed probability distribution over all logically possible words. The distribution has $(|\Sigma| + 1)^2$ *parameters.*

Local Summary

1. Strictly 2-Piecewise models have only $(|\Sigma| + 1)^2$ parameters, but distinguish $2^{|\Sigma|}$ states!

   $$\text{These are, for all } a, b \in \Sigma \cup \{\#\},\ Pr(a \mid b <)$$

2. Under the working hypothesis that likelihood is the same as well-formedness (Coleman and Pierrehumbert 1997, Hayes and Wilson 2008), these parameters can be thought of as independent constraints:

   $$^{*}b \ldots a$$

3. $Pr(\text{s} \mid \text{ʃ,t,o,y,w,n,a} <)$ is **not** independent of $Pr(\text{s} \mid a <)$ nor $Pr(\text{s} \mid \text{ʃ} <)$, etc.
4. This captures the intuition that $Pr(\text{s} \mid \text{ʃ,t,o,y,w,n,a} <)$ is impossibly low because $Pr(\text{s} \mid \text{ʃ} <)$ is!
5. It remains to be shown how to estimate the parameters.

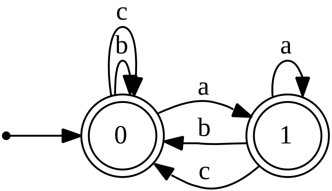
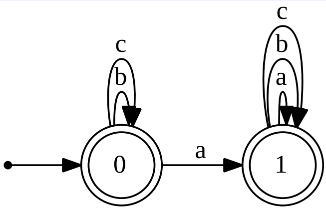
Outline

Section Outline

Estimating SP Distributions

1. Estimating regular distributions
2. Factored models
3. Estimating regular distributions from factors
4. Estimating SP distribution (using factored model)

Strictly Local and Strictly Piecewise Models

Strictly 2-Local (e.g. *st)	Strictly 2-Piecewise (e.g. *s. . . ʃ)
Contiguous subsequences	Subsequences (discontiguous OK)
Immediate Predecessor	Predecessor
Concatenation (·)	Less than (<)
	
0 = have not just seen an [a]	0 = have never seen an [a]
1 = have just seen an [a]	1 = have seen an [a] earlier

(McNaughton and Papert 1971, Simon 1975, Rogers and Pullum 2007, Rogers et. al. 2009, Heinz and Rogers in prep)

Estimating regular distributions
Example with Strictly 2-Local (bigram) model

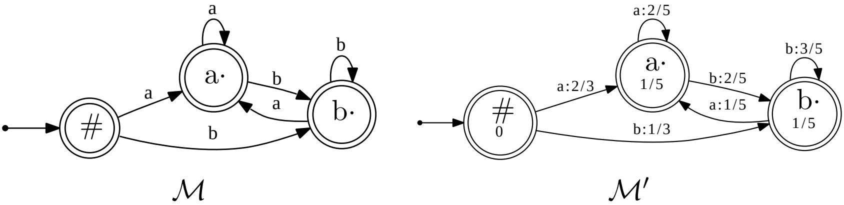


Figure: At left a deterministic finite state acceptors (DFA) representing the family of SL_2 distributions (i.e. bigram model) with $\Sigma = \{a, b\}$. At right, a DFA describing a particular SL_2 distribution.

Theorem (1)

Let $\mathcal{M}$ and $\mathcal{M}'$ be DFAs with the same structure and let $\mathcal{D}_{\mathcal{M}'}$ generate a sample S. Then ***the maximum-likelihood estimate*** *(MLE) of S with respect to $\mathcal{M}$ guarantees that $\mathcal{D}_{\mathcal{M}}$ approaches $\mathcal{D}_{\mathcal{M}'}$ as the size of S goes to infinity.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Estimating regular distributions
Example with Strictly 2-Local (bigram) model

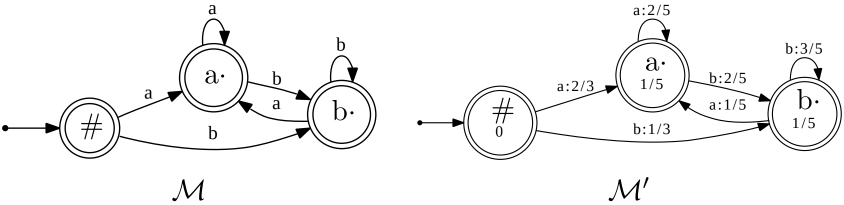


Figure: At left a deterministic finite state acceptors (DFA) representing the family of SL_2 distributions (i.e. bigram model) with $\Sigma = \{a, b\}$. At right, a DFA describing a particular SL_2 distribution.

Theorem (2)

*For a sample S and deterministic finite-state acceptor $\mathcal{M}$, **counting the parse of S through $\mathcal{M}$ and normalizing at each state** optimizes the maximum-likelihood estimate.*

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Estimating regular distributions
Example with Strictly 2-Local (bigram) model

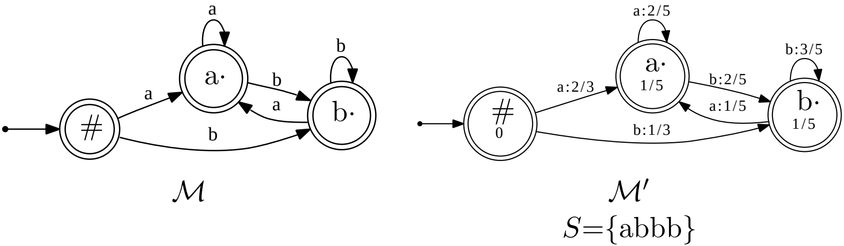
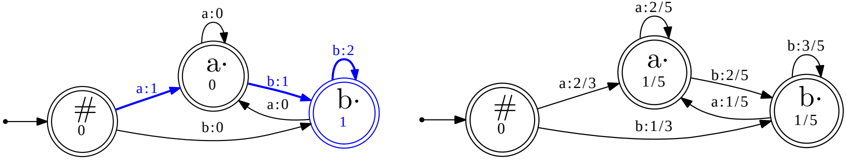


Figure: At left a deterministic finite state acceptors (DFA) representing the family of SL_2 distributions (i.e. bigram model) with $\Sigma = \{a, b\}$. At right, a DFA describing a particular SL_2 distribution.

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Estimating regular distributions Example with Strictly 2-Local (bigram) model



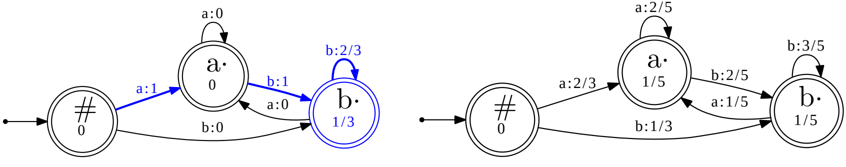
$\mathcal{M}$ $\mathcal{M}'$

S={abbb}

Figure: At left a deterministic finite state acceptors (DFA) representing the family of SL_2 distributions (i.e. bigram model) with $\Sigma = \{a, b\}$. At right, a DFA describing a particular SL_2 distribution.

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

Estimating regular distributions
Example with Strictly 2-Local (bigram) model



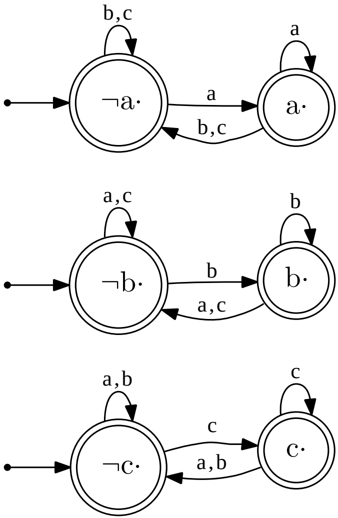
$\mathcal{M}$ $\mathcal{M}'$

S={abbb}

Figure: At left a deterministic finite state acceptors (DFA) representing the family of SL_2 distributions (i.e. bigram model) with $\Sigma = \{a, b\}$. At right, a DFA describing a particular SL_2 distribution.

(Vidal et. al 2005a, 2005b, de la Higuera 2010)

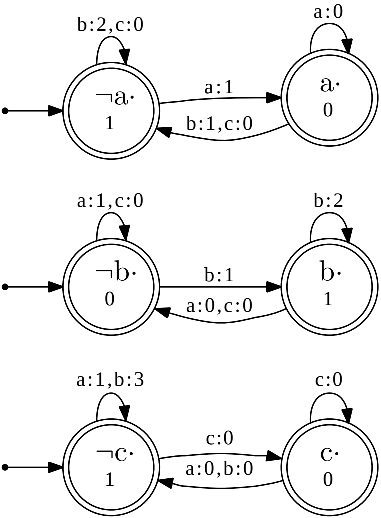
Estimating factored regular distributions



S={abbb}.

A list of DFAs whose product representing the family of strictly 2-local distributions (i.e. bigram model) with $\Sigma = \{a, b, c\}$.

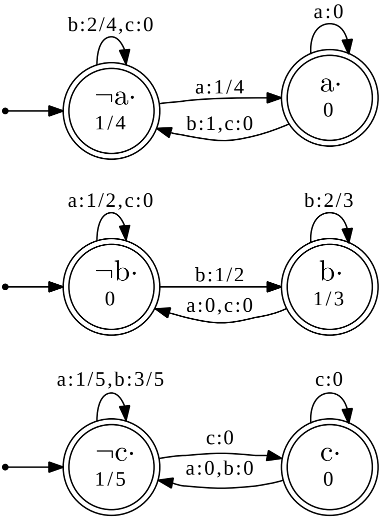
Estimating factored regular distributions



$S = \{\text{abbb}\}$.

A list of DFAs whose product representing the family of strictly 2-local distributions (i.e. bigram model) with $\Sigma = \{a, b, c\}$.

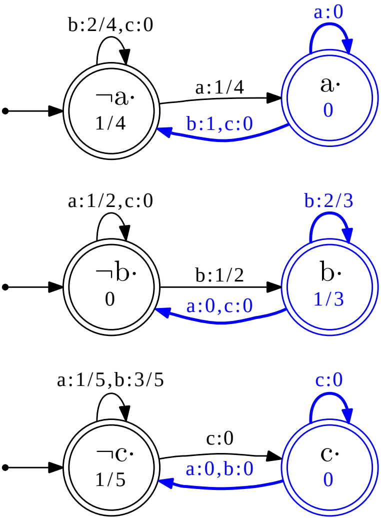
Estimating factored regular distributions



S={abbb}.

A list of DFAs whose product representing the family of strictly 2-local distributions (i.e. bigram model) with $\Sigma = \{a, b, c\}$.

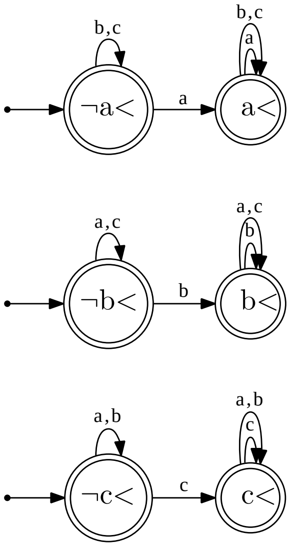
Estimating factored regular distributions



S={abbb}.

A list of DFAs whose product representing the family of strictly 2-local distributions (i.e. bigram model) with $\Sigma = \{a, b, c\}$.

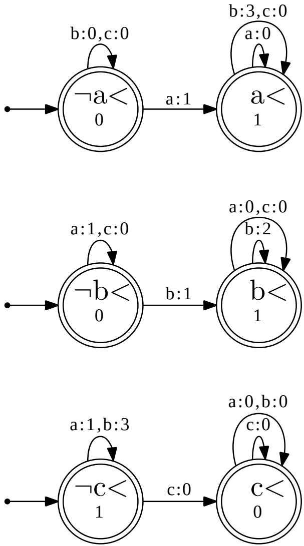
Estimating SP_2 distributions



S={abbb}.

A list of DFAs whose product represents the family of strictly 2-piecewise distributions with $\Sigma = \{a, b, c\}$.

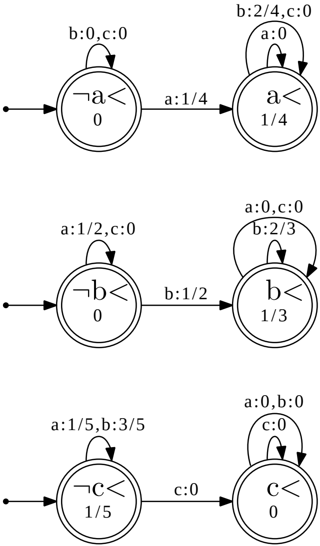
Estimating SP_2 distributions



S={abbb}.

A list of DFAs whose product represents the family of strictly 2-piecewise distributions with $\Sigma = \{a, b, c\}$.

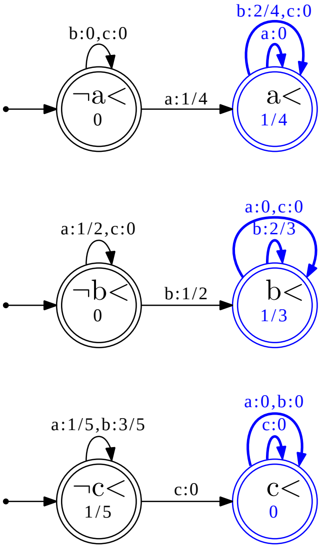
Estimating SP_2 distributions



S={abbb}.

A list of DFAs whose product represents the family of strictly 2-piecewise distributions with $\Sigma = \{a, b, c\}$.

Estimating SP_2 distributions



S={abbb}.

A list of DFAs whose product represents the family of strictly 2-piecewise distributions with $\Sigma = \{a, b, c\}$.

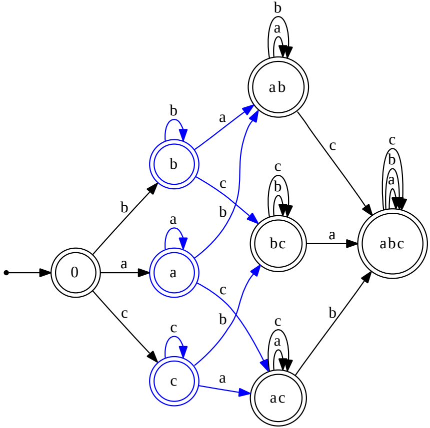
Theorem

This procedure yields the ML estimate for SP distributions.

(Heinz and Rogers, in prep)

Local Summary: Estimating SP_2 Distributions

The automata product of the machines above yields the one below with $2^{|\Sigma|}$ states! But SP distributions only have the parameters on the singleton states; the rest is predictable from Equation 2.



What happens at the initial state can be determined in a variety of fashions.

Outline

Long-distance dependencies

SP Distributions

Estimating SP Distributions

Demo

Issues

Demos

1. Sibilant Harmony in Samala (Ineseño Chumash)
2. Finnish Vowel Harmony
3. English laterals

Samala Corpus

- 4800 words drawn from Applegate 2007, generously provided in electronic form by Applegate (p.c).

35 Consonants

	labial	coronal	a.palatal	velar	uvular	glottal
stop	p pʔ pʰ	t tʔ tʰ		k kʔ kʰ	q qʔ qʰ	ʔ
affricates		t͡s t͡sʔ t͡sʰ	t͡ʃ t͡ʃʔ t͡ʃʰ			
fricatives		s sʔ sʰ	ʃ ʃʔ ʃʰ	x xʔ		h
nasal	m	n nʔ				
lateral		l lʔ				
approx.	w	y				

6 Vowels

i	ɨ	u
e		o
	a	

(Applegate 1972, 2007)

Samala: Results of SP2 estimation

$P(x \mid y <)$		x			
		s	t͡s	ʃ	t͡ʃ
y	s	0.0325	0.0051	0.0013	0.0002
	t͡s	0.0212	0.0114	0.0008	0.
	ʃ	0.0011	0.	0.067	0.0359
	t͡ʃ	0.0006	0.	0.0458	0.0314

(Collapsing laryngeal distinctions)

Finnish: Corpus

- 44,040 words from Goldsmith and Riggle (to appear)

19 Consonants

	lab.	lab.dental	cor.	pal.	velar	uvular	glottal
stop	p b		t d	c	k g	q	
fricatives		f v	s		x		h
nasal	m		n				
lateral			l				
rhotic			r				
approx.	w		j				

8 Vowels

-back		+back
i	y	u
e	oe	o
ae		a

Back vowels and front vowels don't mix (except for [i,e], which are transparent).

Finnish: Results of SP2 estimation

$P(x \mid b <)$		x							
		u	o	a	y	oe	ae	i	e
b	u	0.056	0.040	0.118	0.006	0.002	0.007	0.084	0.072
	o	0.046	0.033	0.120	0.005	0.002	0.007	0.110	0.067
	a	0.045	0.031	0.130	0.005	0.002	0.007	0.095	0.060
	y	0.015	0.016	0.038	0.044	0.026	0.066	0.091	0.072
	oe	0.023	0.027	0.058	0.030	0.014	0.053	0.095	0.067
	ae	0.014	0.014	0.034	0.036	0.015	0.086	0.091	0.073
	i	0.030	0.031	0.097	0.011	0.006	0.0240	0.088	0.080
	e	0.031	0.026	0.077	0.014	0.005	0.031	0.089	0.071

English: CMU Dictionary

1. Carnegie Mellon University American English Pronouncing Dictionary
2. 129,463 words

$P(x \mid y <)$		x: l	x: r
y	l	0.014	0.0535
	r	0.0364	0.0343

(Rhoticized schwa and [r] have been collapsed)

English: Top Unlikely Pairs

1.	CH DH	0.	12.	AY ZH	0.00011
2.	DH CH	0.	13.	SH DH	0.00012
3.	DH DH	0.	14.	W ZH	0.00014
4.	DH ZH	0.	15.	NG OY	0.00014
5.	NG ZH	0.	16.	HH ZH	0.00016
6.	OY OY	0.	17.	Z DH	0.00017
7.	TH DH	0.	18.	Z OY	0.00019
8.	ZH CH	0.	19.	IY DH	0.00019
9.	ZH DH	0.	20.	K DH	0.00021
10.	ZH HH	0.	21.	UH DH	0.00023
11.	ZH TH	0.	22.	UH OY	0.00023

Outline

Typology

Q1: Are there long-distance phonotactic patterns in natural language not licensed by similarity?

Q2: Are the top unlikely pairs accidental generalizations or bonafide internalized generalizations of native speakers?

A: Presumably answerable by artificial language-learning experiments.

Similarity

Q3: Assuming long-distance phonotactic patterns are licensed by similarity, aren't SP distributions too unrestrictive?

A: No. Similarity is an independent filter or bias.

- The role of similarity can now be studied separately.
- It is straightforward to add similarity biases to the model with Bayesian priors

Strictly Local Patterns

Q4: Strictly Piecewise models cannot learn adjacency patterns, right? Aren't they then too restrictive?

A: Right they can't, but No they're not. The perspective here is that phonological learning is modularized:

- One sublearner picks out adjacency patterns (Strictly Local)
- One sublearner picks out long-distance patterns (Strictly Piecewise)
- Plausibly a sublearner for stress patterns (Heinz 2009, Bergelson et. al 2010)
- Finds common ground with a biological perspective (Gallistel and King 2009:Chap 13)

Features

Q5: The SP model looks interesting, but doesn't make use of phonological features. Isn't that a problem?

A: No.

- See poster tomorrow "Feature-Based Generalization" (Heinz and Koirala)
- Also, the SP model can be plugged into the Hayes and Wilson (2008) maxent learner adopting the same featural strategy they apply to Strictly 2-Local constraints (bigrams)

Comparison to Tier-based models

Q6: Since long-distance patterns are learnable by tier-based n-gram models, do we need SP distributions?

A: The models make different predictions, making it a fruitful area for future research.

tier-based SL (n-gram) models	SP models
Captures blocking effects in vowel harmony	Unable to capture blocking effects in vowel harmony
Predicts unattested blocking effects in consonantal harmony	Predicts absence of blocking in consonantal harmony
Only able to describe patterns with transparent vowels if they are "off" the tier	Able to describe patterns with transparent vowels
Requires independent theory of tiers	Does not require independent theory of tiers

Conclusion

Gradient phonotactic long-distance dependencies are provably feasibly learnable by estimating SP distributions

This happens

1. without tiers
2. without a concept of similarity
3. without the additional structure provided by OT or P&P frameworks
4. with a naturally structured hypothesis space
5. with a learner whose behavior is provably correct

Acknowledgments

Thanks to the University of Delaware Phonology and Phonetics group, the participants in the Fall 2009 seminar on "Likelihood and Well-formedness in Phonology," and especially Karthik Durvasula, Bill Idsardi, Cesar Koirala, and Robert Wilder for valuable discussion.

This research was supported by a 2008-2009 University of Delaware Research Fund Grant and a summer 2009 General University Research grant.