Rethinking representations

A log-bilinear model of phonotactics

Huteng Dai (Rutgers), Connor Mayer (UC Irvine), and Richard Futrell (UC Irvine)

Take-home messages

- A novel representational system: continuous features
- A log-bilinear model compatible with both continuous and discrete features
- Finding: In several cases, models with continuous representations outperformed their counterparts

ROADMAP

1. Phonotactic learning and features
2. A log-bilinear model of phonotactic learning
3. Model/feature comparison
4. Conclusions and future directions

Phonotactics

Restrictions on how sounds can be sequenced;

Phonotactics **vary across languages** and must be **learned**

- /st/ onset is acceptable in English, but not in Spanish

Gradient acceptability in phonotactics

Gradient well-formedness is often found in acceptability experiments. (e.g. Coleman & Pierrehumbert 1997, Albright 2009, Hayes et al. 2009, Daland et al. 2011)

- _**bl**ick_ >> $^{?}$_**bw**ick_ >> *_**bn**ick_ >> **_**bz**ick_ (Albright 2009)

This motivated **probabilistic** models of phonotactics

(Hayes & Wilson 2008, Futrell et al. 2017, Mayer & Nelson 2020; cf. Gorman 2013, Kahng & Durvasula 2023)

Why features?

Segmental generalizations often overlook sub-segmental properties

- b[+approximant] ([bj, br, bl]) is highly frequent
- No b[-approximant]
- This explains why [bw] >> [bn] even though both unattested in English

This motivates sub-segmental representations such as **phonological features**.

Traditional view of phonological features

- **Universal:** all languages described by same set of features
- **Phonetically-based**: reflect phonetic properties
- **Discrete:** values are +, –, or 0

$$/\mathrm{p}/ = \begin{bmatrix} +\text{LABIAL} \\ -\text{continuant} \\ -\text{voice} \\ \vdots \end{bmatrix}$$

Traditional view of phonotactic learning

- Input: training data (lexicon) + universal feature system

1. Training data

berari
boka
pupabopa
pabarubo
…

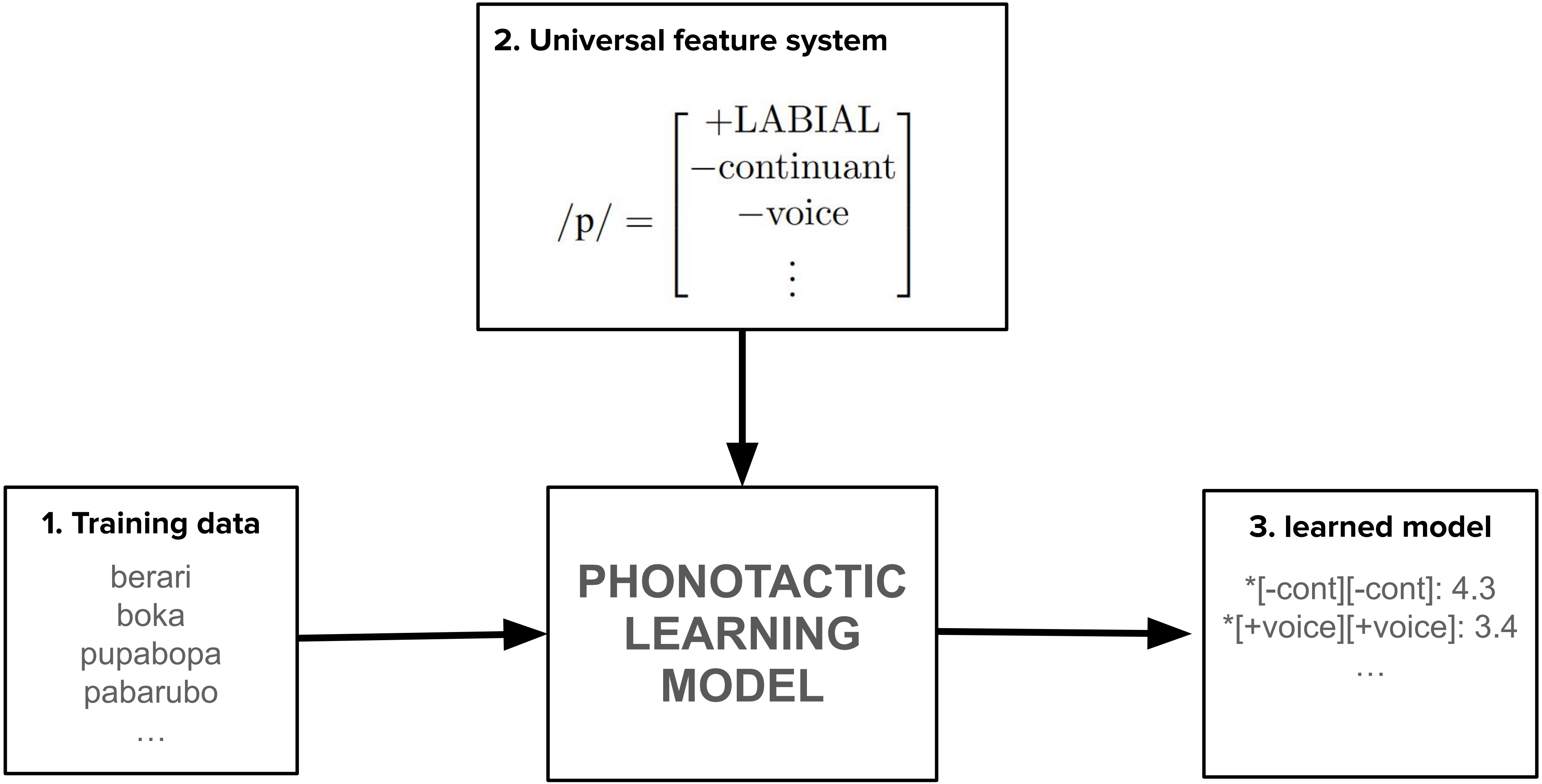
2. Universal feature system

$$/\mathrm{p}/ = \begin{bmatrix} +\mathrm{LABIAL} \\ -\mathrm{continuant} \\ -\mathrm{voice} \\ \vdots \end{bmatrix}$$

- Output: learned model
- The learning succeeds if the learned model predicts a probabilistic distribution that matches the acceptability of nonce forms.

(Hayes & Wilson 2008)

Traditional phonotactic learning with universal features



Challenge: processes of unnatural classes

Many phonological classes don't share phonetic properties. (Mielke 2008)

(2) Evenki post-nasal nasalization (Mielke 2008; Nedjalkov 1997)

i. Evenki productive suffixation

a.	/oron-vi/	oronmi	'my reindeer'
b.	/ŋinakin-si/	ŋinakinni	'your dog'
c.	/oron-gAtʃin/	oronŋotʃin	'like a reindeer'
Cf.			
d.	/amkin-du/	amkindu	'bed-DATIVE'
e.	/ekun-da/	ekunda	'somebody, something'

ii. Evenki nasalization
{v, s, g} → {m, n, ŋ}/ [+nasal]___

Invent a new universal feature for every unnatural class?

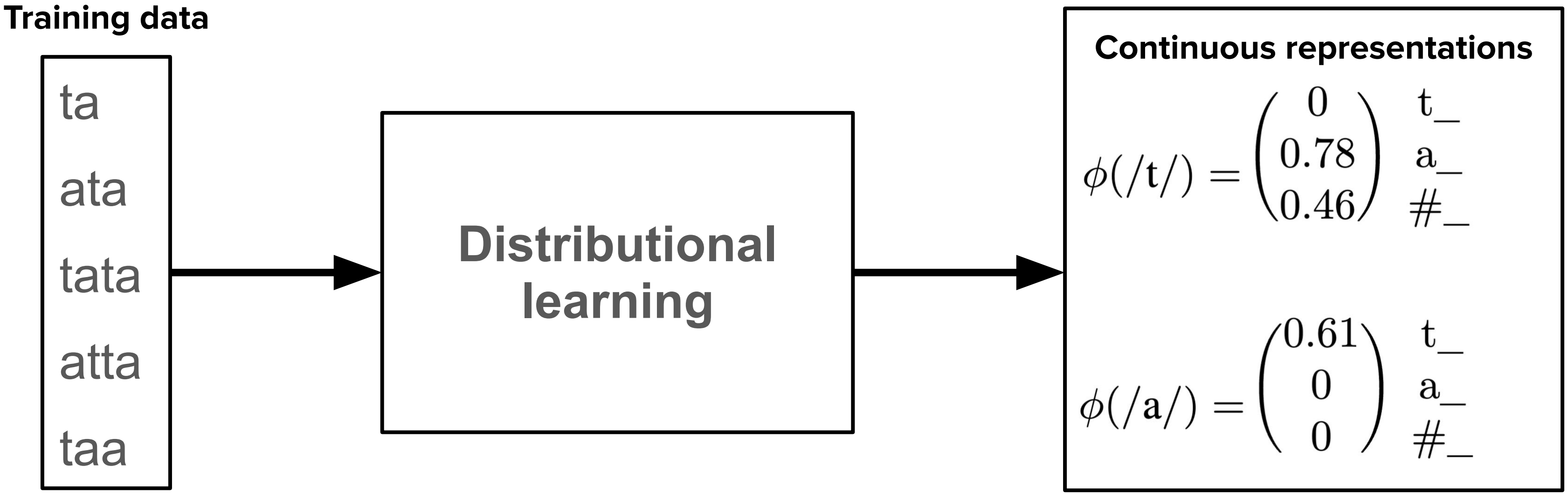
Our “Emergent” view of features

- **Language-specific**
- **Learned or emergent**
- **Distributional:** shared contexts (e.g. {v, s, g}/[-nasal]_) implies shared features;

(Mielke 2008, Nazarov 2014, 2016, Archangeli & Pulleyblank 2018, 2022, Gallagher 2019)

Distributional learning: continuous representations

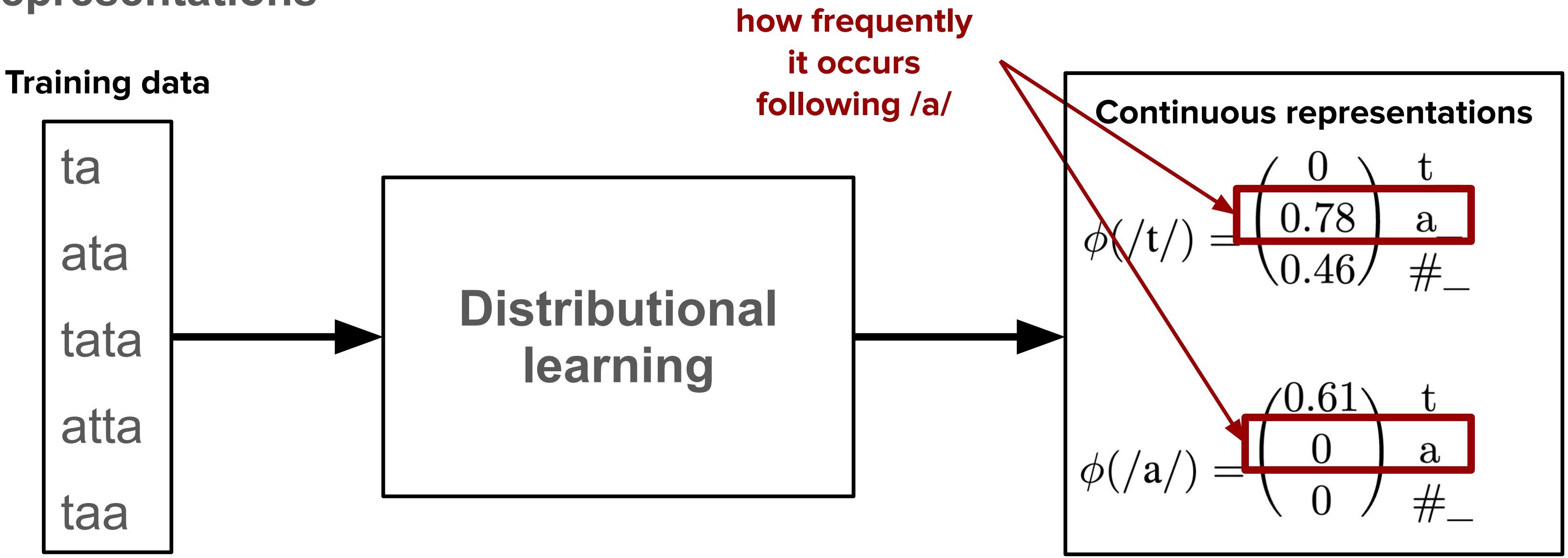
Distributional learning models produce **continuous (real-valued) representations**



(e.g. Goldsmith & Xanthos 2009, Mayer 2020, Nelson 2022, a.o.)

Distributional learning: continuous representations

Distributional learning models produce **continuous (real-valued) representations**



(e.g. Goldsmith & Xanthos 2009, Mayer 2020, Nelson 2022, a.o.)

Distributional learning: discretization

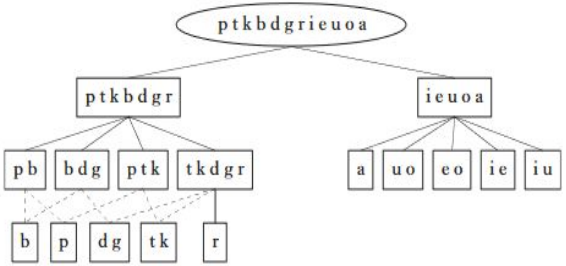
1. Clustering to produce classes (Goldsmith & Xanthos 2009, Mayer 2020)
2. Derive feature system from sets of classes (Mayer & Daland 2020)

Continuous representations

$$\phi(/\mathrm{t}/) = \begin{array}{c} \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} \end{array} \begin{array}{l} \mathrm{t\_} \\ \mathrm{a\_} \\ \#\_ \end{array}$$

$$\phi(/\mathrm{a}/) = \begin{array}{c} \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} \end{array} \begin{array}{l} \mathrm{t\_} \\ \mathrm{a\_} \\ \#\_ \end{array}$$

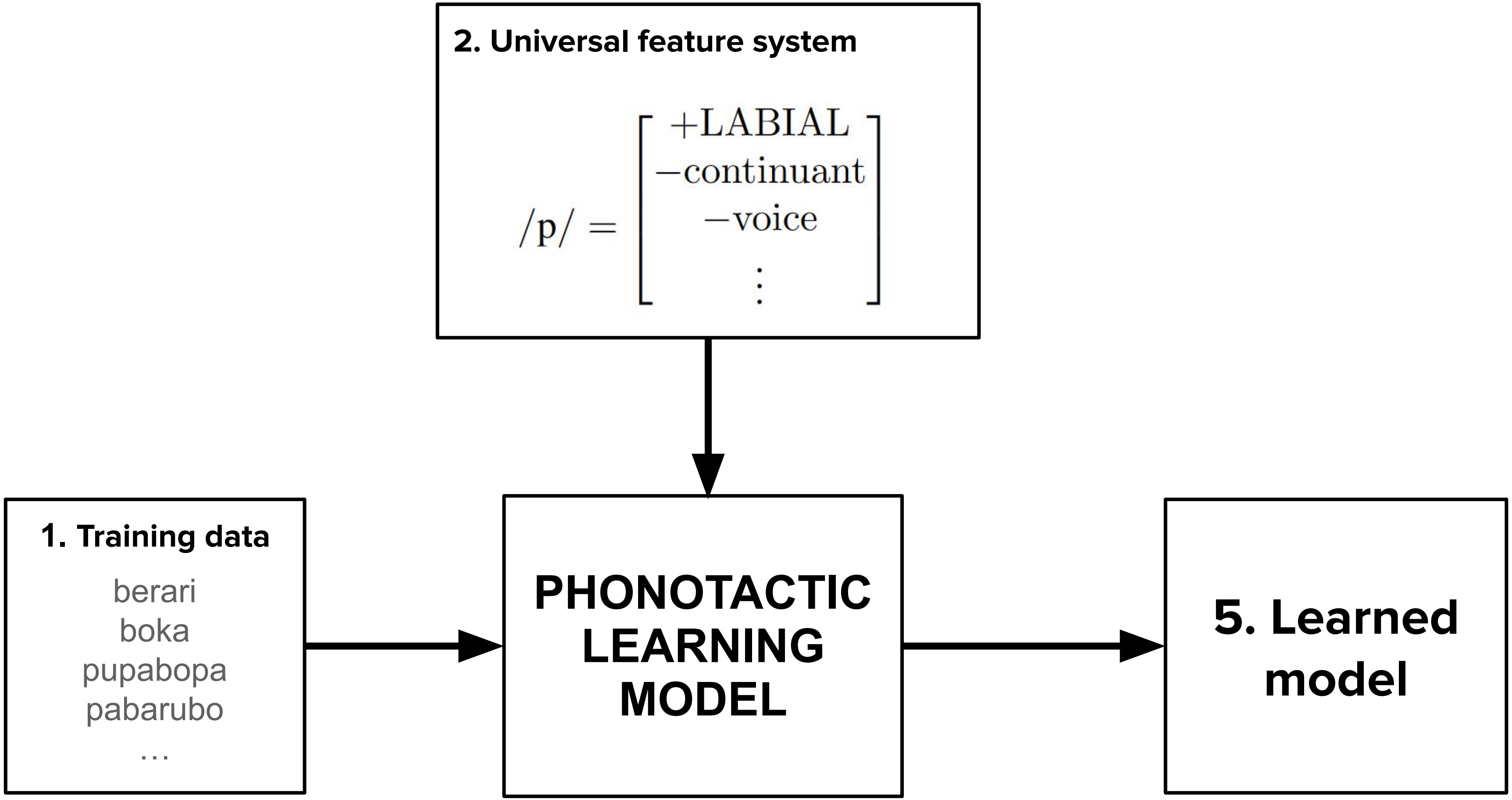
Induced classes



Derived discrete feature

$$/\mathrm{p}/ = \begin{bmatrix} -\mathrm{f1} \\ +\mathrm{f2} \\ -\mathrm{f3} \\ \vdots \end{bmatrix}$$

Traditional phonotactic learning with universal features



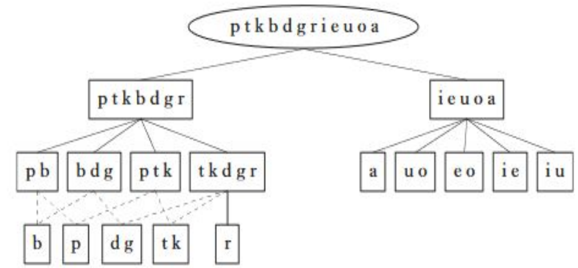
Phonotactic learning with derived discrete features

2. Continuous representations

$$\phi(/\text{t}/) = \begin{array}{c} \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} \end{array} \begin{array}{l} \text{t\_} \\ \text{a\_} \\ \#\text{\_} \end{array}$$

$$\phi(/\text{a}/) = \begin{array}{c} \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} \end{array} \begin{array}{l} \text{t\_} \\ \text{a\_} \\ \#\text{\_} \end{array}$$

3. Induced classes



4. Derived discrete feature

$$/\text{p}/ = \begin{bmatrix} -\text{f1} \\ +\text{f2} \\ -\text{f3} \\ \vdots \end{bmatrix}$$

1. Training data

berari
boka
pupabopa
pabarubo
…

PHONOTACTIC LEARNING MODEL

5. Learned model

Correlation with phonetic distinctions

Learned distributional representations can reflect phonetic distinctions;

(Goldsmith & Xanthos 2009, Mayer 2020)

Perform comparably to phonetic features in phonotactic learning

(Nelson 2022)

Challenge from discretization

- Too many steps
- Some information from continuous representations is discarded

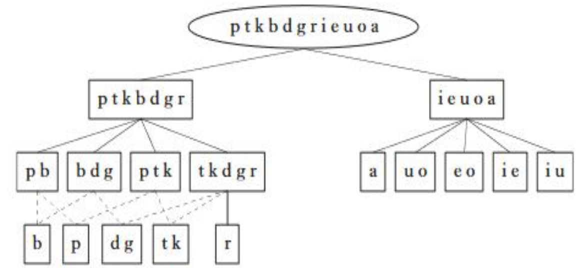
Phonotactic learning with derived discrete features

2. Continuous representations

$$\phi(/\mathrm{t}/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} \begin{array}{l} \mathrm{t}\_ \\ \mathrm{a}\_ \\ \#\_ \end{array}$$

$$\phi(/\mathrm{a}/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} \begin{array}{l} \mathrm{t}\_ \\ \mathrm{a}\_ \\ \#\_ \end{array}$$

3. Induced classes



4. Derived discrete feature

$$/\mathrm{p}/ = \begin{bmatrix} -\mathrm{f1} \\ +\mathrm{f2} \\ -\mathrm{f3} \\ \vdots \end{bmatrix}$$

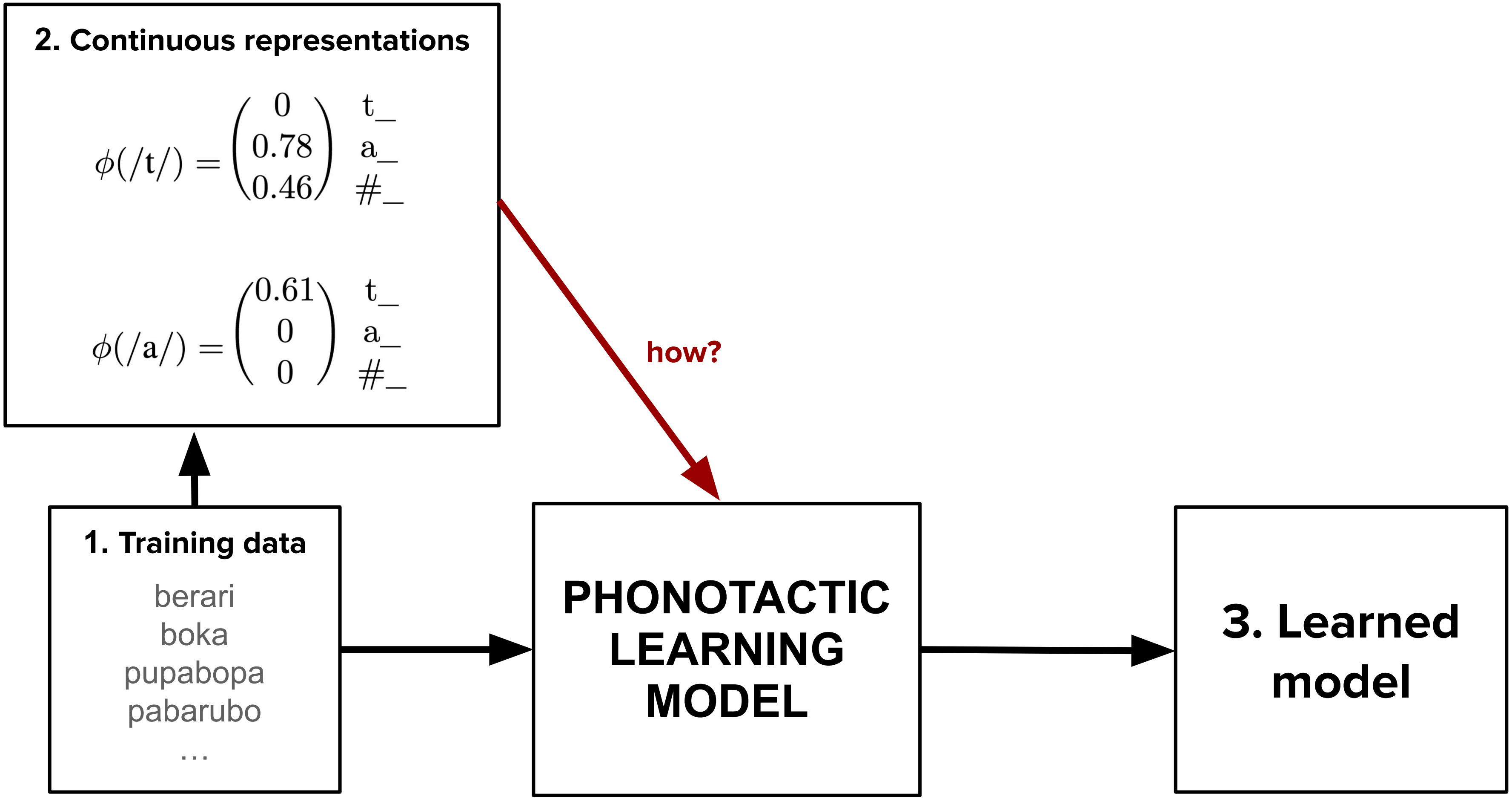
1. Training data

berari
boka
pupabopa
pabarubo
…

PHONOTACTIC LEARNING MODEL

5. Learned model

Another possibility



ROADMAP

1. Phonotactic learning and features
2. A log-bilinear model of phonotactic learning (20 slides left!)
3. Model/feature comparison
4. Conclusions and future directions

A log-linear model

In a **log-linear** (Maximum Entropy) model, the probability of an outcome *x* is

$$p(x) \propto \exp\left\{\mathbf{w}^\top \phi(x)\right\}$$

A log-linear model

In a **log-linear** (Maximum Entropy) model, the probability of an outcome *x* is

$$p(x) \propto \exp\left\{\mathbf{w}^\top \phi(x)\right\}$$

$$\mathbf{w}^\top \phi(x) = w_1 \phi_1(x) + w_2 \phi_2(x) + \cdots$$

weight for first feature

value of first feature

A log-linear model

In a **log-linear** (Maximum Entropy) model, the probability of a outcome *x* is

$$p(x) \propto \exp\left\{\underbrace{\mathbf{w}}_{\text{constraint weights}}{}^{\top}\underbrace{\phi(x)}_{\text{constraint violations by form } x}\right\}$$

In Hayes & Wilson (2008): constraint weights

constraint violations by form *x*

ϕ: learned or engineered

$\mathbf{w}$: learned from data

A log-linear model

In a **log-linear** (Maximum Entropy) model, the probability of a outcome *x* is

$$p(x) \propto \exp\left\{\mathbf{w}^\top \phi(x)\right\}$$

How can we make this *conditional*, so we can calculate the *probability of a segment given context?* e.g. $p(\mathrm{bl}) = p(\mathrm{b} \mid \#) \cdot p(\mathrm{l} \mid \#\mathrm{b})$

$$p(x \mid c) = \ ?$$

a **segment**

a **context**

A log-***bi***linear model: overview

In a **log-bilinear** model, the probability of a segment *x* given context *c* is

$$p(x \mid c) \propto \exp\left\{\underline{\psi(c)}^\top \underline{\mathbf{A}}\underline{\phi(x)}\right\}$$

In our model: **feature vector of context *c*** **feature vector of segment *x***

A guides how to connect the features of *c* and *x*

A log-***bi***linear model: interaction matrix A

In a **log-bilinear** model, the probability of a segment *x* given context *c* is

$$p(x \mid c) \propto \exp\left\{\psi(c)^\top \mathbf{A}\phi(x)\right\}$$

Weight matrix $\mathbf{A_{ij}}$: how likely a feature $\phi_i(x)$ co-occur with feature $\psi_j(c)$.

$$\begin{array}{cc} & \begin{array}{ccc} \psi_1(c) & \psi_2(c) & \psi_3(c) \end{array} \\ \begin{array}{c} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \end{array} & \begin{pmatrix} -0.174 & 0.152 & 0.314 \\ 0.118 & -0.011 & 0.236 \\ 0.530 & 0.512 & -0.861 \end{pmatrix} \end{array}$$

A is learned by gradient descent to maximize likelihood of training data.

A log-***bi***linear model: interaction matrix A

In a **log-bilinear** model, the probability of a segment *x* given context *c* is

$$p(x \mid c) \propto \exp\left\{\psi(c)^\top \mathbf{A}\phi(x)\right\}$$

$$\psi(c)^\top \mathbf{A}\phi(x) = a_{11}\psi_1(c)\phi_1(x) + a_{12}\psi_1(c)\phi_2(x) + \ldots a_{21}\psi_2(c)\phi_1(x) + \ldots$$

weight for first context feature and first segment feature

value of first context feature

value of first segment feature

ROADMAP

1. Phonotactic learning and features
2. A log-bilinear model of phonotactic learning
3. Model/feature comparison (15 slides left!)
4. Conclusions and future directions

Compatibility

Log-bilinear model is compatible to all types of featural representations;

We test the model using **three types of featural representations**

1. Discrete phonetic features
2. Continuous distributional features
3. Discretized distributional features

Type 1: Discrete phonetic features

We use the feature specifications from Hayes (2009)

- Segment is either 1 or 0 for each feature-value pair

$$\phi(\mathrm{k}) = \begin{array}{cl} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \end{bmatrix} & \begin{array}{c} +\mathrm{dorsal} \\ -\mathrm{dorsal} \\ +\mathrm{continuous} \\ -\mathrm{continuous} \\ +\mathrm{consonantal} \\ -\mathrm{consonantal} \\ \vdots \end{array} \end{array}$$

Type 1: Discrete phonetic features

Hayes (2009)

2. Universal feature system

$$\phi(\mathrm{k}) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \begin{matrix} +\text{dorsal} \\ -\text{dorsal} \\ +\text{continuous} \\ -\text{continuous} \\ +\text{consonantal} \\ -\text{consonantal} \\ \vdots \end{matrix}$$

1. Training data

berari
boka
pupabopa
pabarubo
…

log-bilinear model

3. Learned model

$$\begin{pmatrix} -0.174 & 0.152 & 0.314 \\ 0.118 & -0.011 & 0.236 \\ 0.530 & 0.512 & -0.861 \end{pmatrix}$$

Type 2: Continuous distributional features

Dimensions: preceding and following bigram contexts (Mayer 2020)

Values: Calculated in two steps

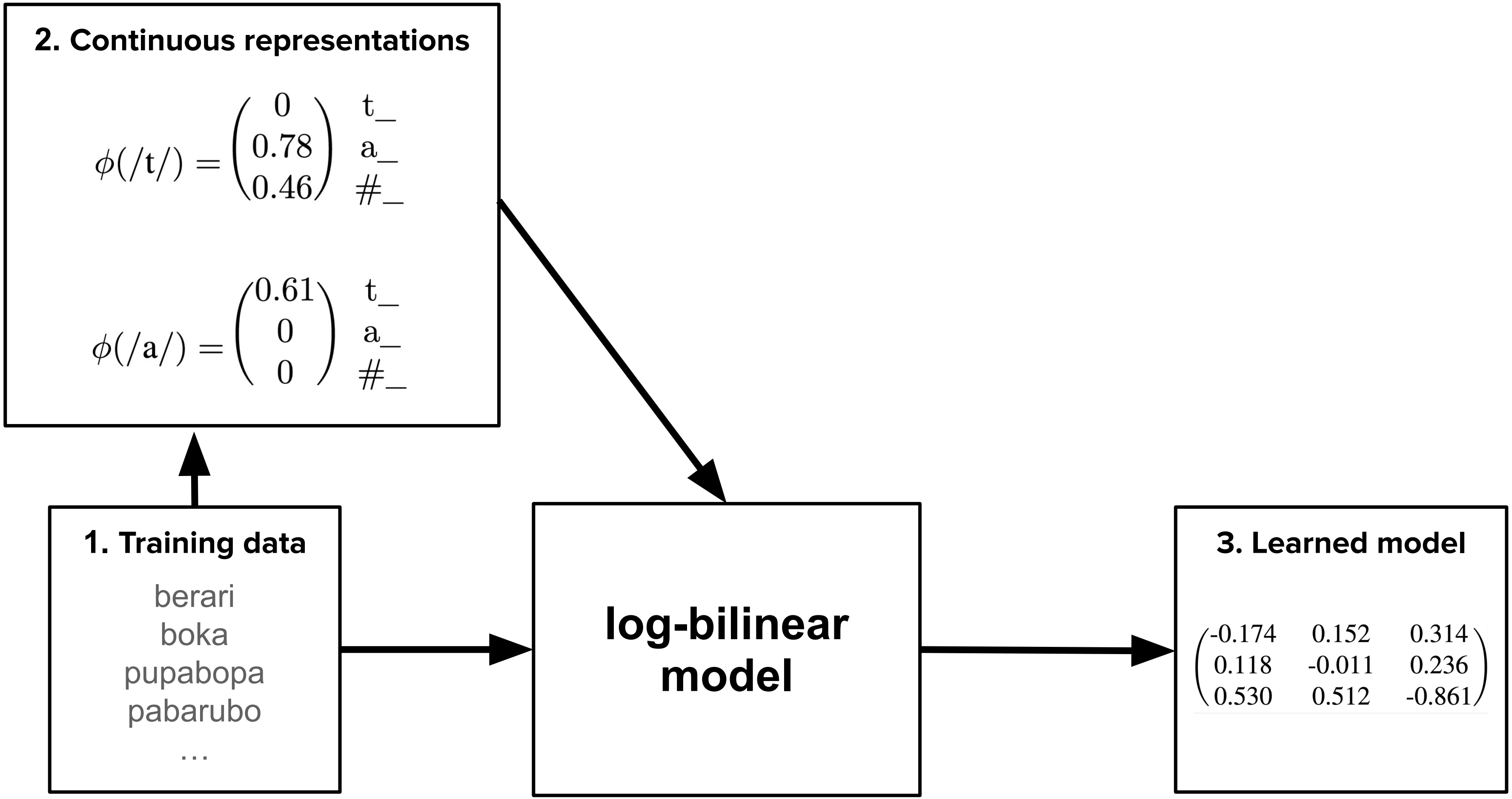
1. Compute **bigram probabilities** using a smoothed bigram language model
2. Convert probabilities to **Pointwise mutual information (PMI):**

$$\phi(/\mathrm{t}/) = \begin{array}{cc} \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} & \begin{array}{c} \mathrm{t}\_ \\ \mathrm{a}\_ \\ \#\_ \end{array} \end{array}$$

$$\mathrm{PMI}(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

$$\phi(/\mathrm{a}/) = \begin{array}{cc} \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} & \begin{array}{c} \mathrm{t}\_ \\ \mathrm{a}\_ \\ \#\_ \end{array} \end{array}$$

Type 2: Continuous distributional features

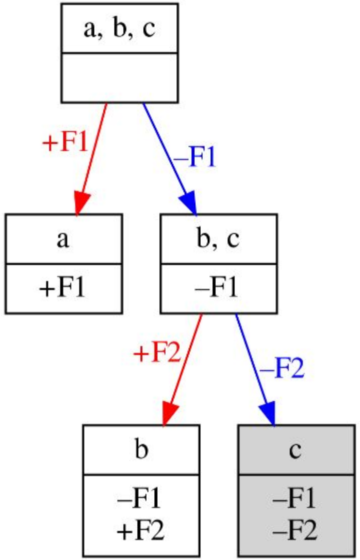


Type 3: discretized distributional features

Starting point: continuous distributional features

Steps:

1. Run clustering algorithm from Mayer (2020) to convert into classes
2. Run algorithm from to derive feature system that describes classes (Mayer & Daland 2020)



σ	F1	F2
a	+	0
b	-	+
c	-	-

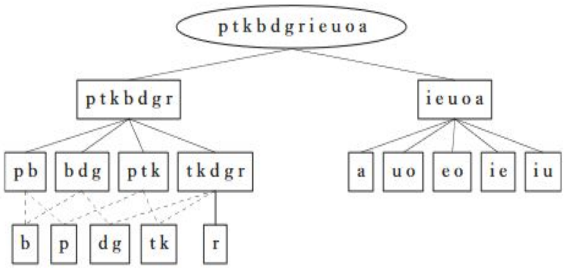
Type 3: discretized distributional features

2. Continuous representations

$$\phi(/\mathrm{t}/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} \begin{array}{l} \mathrm{t}\_ \\ \mathrm{a}\_ \\ \#\_ \end{array}$$

$$\phi(/\mathrm{a}/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} \begin{array}{l} \mathrm{t}\_ \\ \mathrm{a}\_ \\ \#\_ \end{array}$$

3. Induced classes



4. Derived discrete feature

$$/\mathrm{p}/ = \begin{bmatrix} -\mathrm{f1} \\ +\mathrm{f2} \\ -\mathrm{f3} \\ \vdots \end{bmatrix}$$

1. Training data

berari
boka
pupabopa
pabarubo
…

log-bilinear model

5. Learned model

$$\begin{pmatrix} -0.174 & 0.152 & 0.314 \\ 0.118 & -0.011 & 0.236 \\ 0.530 & 0.512 & -0.861 \end{pmatrix}$$

Testing the models and featurizations on English onsets

Training data: English onset corpus from Hayes & Wilson (2008)

- 31,641 unlabelled onsets from CMU Pronouncing Dictionary (Weide et al. 1998)

Testing data: Experimental data from Daland et al. (2011)

- Likert ratings given to English nonce words with 48 different onsets by 48 participants
- Broken down into attested, marginal (type frequency < 11), and unattested

Model comparison

We also compare it against **three other phonotactic learning models:**

- Benchmark: Hayes & Wilson learner (Hayes & Wilson 2008)
- MaxEntGrams (Nelson 2022)
- Smoothed bigram model

Model comparison

We also compare it against **three other phonotactic learning models:**

- Benchmark: Hayes & Wilson learner (Hayes & Wilson 2008)
- MaxEntGrams (Nelson 2022)
- Smoothed bigram model

See final paper for these results

Training procedure

Log-bilinear model

- All three types of features
- Cross-validation done to select optimal hyperparameters

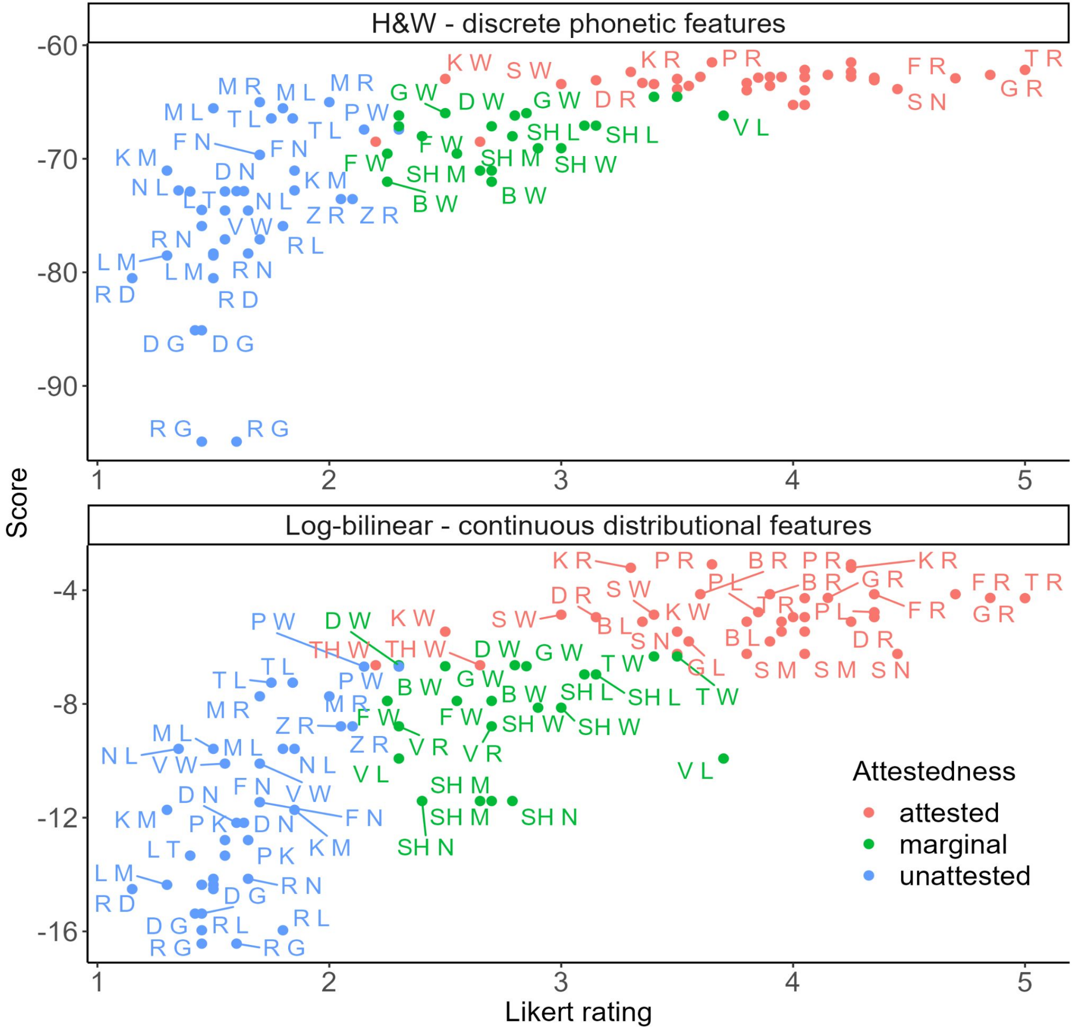
Hayes & Wilson learner (Benchmark)

- Discrete phonetic features and discretized distributional feature
- Maximum of 300 constraints
- Default O/E threshold of 0.3

Result: Kendall's τ correlation

Model	Featurization	Overall	Attested	Marginal	Unattested
H&W	discrete phon.	0.674	0.261	**0.301**	0.374
	discrete dist.	0.634	0.244	-0.049	0.421
Bilinear	discrete phon.	0.646	0.215	0.247	0.377
	discrete dist.	0.572	0.296	0.067	0.309
	continuous dist.	**0.694**	**0.332**	0.201	**0.465**

Comparing the two best models



Future directions

New data and new patterns

- We found our model inherently predicts distance decay (Zymet 2015)

Fine-grained phonetic features (Mielke 2012)

The definition of ‘context’ is flexible

- We focus local context
- Could be be extended to different types of contexts

Conclusion

- ❖ A log-bilinear model compatible with both continuous and discrete features
- ❖ A technique of learning featural representations from the distribution
- ❖ Finding: In several cases, models with continuous representations outperformed their counterparts

Thank you!

Q & A

Discussion

The log-bilinear model with continuous features outperforms the same model with discretized features

- We lose relevant information when we discretize them

Model	Featurization	Overall		Attested		Marginal		Unattested	
		r	τ	r	τ	r	τ	r	τ
Smoothed bigram	segments	**0.877**	0.669	0.509	0.244	0.274	-0.004	0.470	0.280
MaxEntGrams	discrete dist.	0.753	0.610	0.424	0.282	0.212	0.171	**0.583**	0.417
H&W	discrete phon.	0.740	0.674	0.533	0.261	**0.422**	**0.301**	0.459	0.374
	discrete dist.	0.818	0.634	0.540	0.244	-0.012	-0.049	0.547	0.421
Bilinear	discrete phon.	0.785	0.646	0.446	0.215	0.367	0.247	0.525	0.377
	discrete dist.	0.757	0.572	0.520	0.296	0.021	0.067	0.523	0.309
	continuous dist.	0.699	**0.694**	**0.611**	**0.332**	0.247	0.201	0.562	**0.465**

Table 5: Model comparison using Pearson's r and Kendall's τ to correlate model scores with acceptability ratings for English onsets. The correlation value for the top performing model in each category is bolded.

Results

Model	Featurization	Overall		Attested		Marginal		Unattested	
		r	τ	r	τ	r	τ	r	τ
H&W	discrete phon.	0.740	0.674	0.533	0.261	**0.422**	**0.301**	0.459	0.374
	discrete dist.	0.818	0.634	0.540	0.244	-0.012	-0.049	0.547	0.421
Bilinear	discrete phon.	0.785	0.646	0.446	0.215	0.367	0.247	0.525	0.377
	discrete dist.	0.757	0.572	0.520	0.296	0.021	0.067	0.523	0.309
	continuous dist.	0.699	**0.694**	**0.611**	**0.332**	0.247	0.201	0.562	**0.465**

r = Pearson's rho τ = Kendall's tau

What are “features”?

Usually: A **discrete** representational system we used to rationalize the internal structure of basic linguistic representation, such as phonemes.

Some of them have phonetic underpinnings. However, the space of phonetic representations itself is a continuum. e.g. i—e.

Most previous phonotactic models require a prespecified feature file with segments corresponding values in discrete features.

What are “features”?

But also:

We can learn **continuous** representations from distributions: they function just as well as discrete representation, see Mayer (2020).

Proposals for continuous phonetic features (Mielke, 2012)

=> **How would a phonotactics model work that operates natively over continuous features, without discretizing?**

Two types of research in computational phonology

1. Mathematical underpinning of phonological patterns
2. Modeling human performance

We are the second type

Put discrete featural representation in a matrix

	sonorant	voice	labial
p	-	-	+
b	-	+	+
m	+	+	+

Put discrete featural representation in a matrix

	sonorant	voice	labial
p	0	0	1
b	0	1	1
m	1	1	1

Open question: continuous phonetic feature?

	sonorant	voice	labial
p	0	0	1
b	0	1	1
m	1	1	1

Put discrete featural representation in a matrix

	#_l	#_r	#_n
p	1	1	0
b	1	1	0
m	0	0	0

PMI

	#_l	#_r	#_n
p	2.464	1.934	0
b	2.464	1.934	0
m	0	0	0