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)

- *blick* $>>$ *bwick* $>>$ *bnick* $>>$ **bzick**
  
  (Albright 2009)

This motivated **probabilistic** models of phonotactics

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

\[ /p/ = \begin{bmatrix} + \text{LABIAL} \\ - \text{continuant} \\ - \text{voice} \\ \vdots \end{bmatrix} \]
Traditional view of phonotactic learning

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

1. Training data
   - berari
   - boka
   - pupabopa
   - pabarubo
   - …

2. Universal feature system
   \[ /p/ = \begin{bmatrix} +\text{LABIAL} \\ -\text{continuant} \\ -\text{voice} \\ \vdots \end{bmatrix} \]

- 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

2. Universal feature system

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

1. Training data
   berari
   boka
   pupabopa
   pabarubo
   ...

3. Learned model
   *[-cont][-cont]: 4.3
   * [+voice][+voice]: 3.4
   ...

PHONOTACTIC LEARNING MODEL
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-gAtʃ in/ oronŋotʃ in ‘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;

Distributional learning: continuous representations

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

![Diagram](image)

Training data:
- ta
- ata
- tata
- atta
- taa

Continuous representations:
- $\phi(/t/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \\ \# \end{pmatrix}$
- $\phi(/a/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \\ \# \end{pmatrix}$

(e.g. Goldsmith & Xanthos 2009, Mayer 2020, Nelson 2022, a.o.)
Distributional learning: continuous representations

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

Training data

- ta
- ata
- tata
- atta
- taa

Distributional learning

Continuous representations

\[
\phi(/t/) = \begin{pmatrix} 0.78 \\ 0.46 \end{pmatrix}
\]

\[
\phi(/a/) = \begin{pmatrix} 0.61 \\ 0 \end{pmatrix}
\]

how frequently it occurs following /a/

(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)

\[
\phi(/t/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix}
\]
\[
\phi(/a/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix}
\]

Induced classes

Derived discrete feature

\[
/p/ = \begin{bmatrix} -f_1 \\ +f_2 \\ -f_3 \\ \vdots \end{bmatrix}
\]
Traditional phonotactic learning with universal features

1. Training data
   berari
   boka
   pupabopa
   pabarubo
   ...

2. Universal feature system

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

PHONOTACTIC LEARNING MODEL

5. Learned model
2. Continuous representations

\[ \phi(\text{/t/}) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} \]

\[ \phi(\text{/a/}) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} \]

3. Induced classes

PHONOTACTIC LEARNING MODEL

4. Derived discrete feature

\[ /p/ = \begin{bmatrix} -f_1 \\ +f_2 \\ -f_3 \\ \vdots \end{bmatrix} \]

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

\[
\phi(/t/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} \begin{pmatrix} t_- \\ a_- \\ #_- \end{pmatrix}
\]

\[
\phi(/a/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} t_- \\ a_- \\ #_- \end{pmatrix}
\]

3. Induced classes

4. Derived discrete feature

\[
/p/ = \begin{bmatrix} -f_1 \\ +f_2 \\ -f_3 \\ \vdots \end{bmatrix}
\]

PHONOTACTIC LEARNING MODEL

1. Training data

berari
boka
pupabopa
pabarubo
...

5. Learned model
2. Continuous representations

\[ \phi(/t/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \end{pmatrix} \]

\[ \phi(/a/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \end{pmatrix} \]

Another possibility

1. Training data
   berari
   boka
   pupabopa
   pabarubo
   ...

2. PHONOTACTIC LEARNING MODEL

3. Learned model

how?
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

weight for first feature
A log-linear model

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

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

- $\phi$: learned or engineered
- $\mathbf{w}$: learned from data

In Hayes & Wilson (2008): constraint weights

constraint violations by form $x$
A log-linear model

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

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

How can we make this *conditional*, so we can calculate the probability of a segment given context? e.g. $p(\text{bl}) = p(b \mid \#) \cdot p(1 \mid \#b)$

$$p(x \mid c) = ?$$
A log-bilinear model: overview

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

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

**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-bilinear model: interaction matrix A

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

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

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

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

$A$ is learned by gradient descent to maximize likelihood of training data.
A log-bilinear model: interaction matrix A

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

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

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

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

\[
\phi(k) = \begin{bmatrix}
1 \\
0 \\
0 \\
1 \\
0 \\
\vdots
\end{bmatrix}
\begin{align*}
+\text{dorsal} \\
-\text{dorsal} \\
+\text{continuous} \\
-\text{continuous} \\
+\text{consonantal} \\
-\text{consonantal} \\
\vdots
\end{align*}
\]
Type 1: Discrete phonetic features

2. Universal feature system

\[ \phi(k) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \]

- dorsal
- +dorsal
- -dorsal
- +continuous
- -continuous
- +consonantal
- -consonantal

Hayes (2009)

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)**:

\[
\text{PMI}(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}
\]

\[
\phi(/t/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \\ 0 \end{pmatrix}
\]

\[
\phi(/a/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\]
Type 2: Continuous distributional features

2. Continuous representations

\[
\phi(/t/) = \begin{pmatrix} 0 \\ 0.78 \\ 0.46 \\ \#_\_ \end{pmatrix}
\]

\[
\phi(/a/) = \begin{pmatrix} 0.61 \\ 0 \\ 0 \\ \#_\_ \end{pmatrix}
\]

1. Training data
   berari
   boka
   pupabopa
   pabarubo
   ...

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 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)
Type 3: discretized distributional features

1. Training data
   berari
   boka
   pupabopa
   pabarubo
   ...

2. Continuous representations
   \[
   \phi(/t/) = \begin{pmatrix}
   0 \\
   0.78 \\
   0.46 \\
   \end{pmatrix}, \quad \phi(/a/) = \begin{pmatrix}
   0.61 \\
   0 \\
   0 \\
   \end{pmatrix}
   \]

3. Induced classes

4. Derived discrete feature
   \[/p/ = \begin{bmatrix}
   -f_1 \\
   +f_2 \\
   -f_3 \\
   \vdots \\
   \end{bmatrix}\]

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 $\tau$ correlation

<table>
<thead>
<tr>
<th>Model</th>
<th>Featurization</th>
<th>Overall</th>
<th>Attested</th>
<th>Marginal</th>
<th>Unattested</th>
</tr>
</thead>
<tbody>
<tr>
<td>H&amp;W</td>
<td>discrete phon.</td>
<td>0.674</td>
<td>0.261</td>
<td>0.301</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td>discrete dist.</td>
<td>0.634</td>
<td>0.244</td>
<td>-0.049</td>
<td>0.421</td>
</tr>
<tr>
<td>Bilinear</td>
<td>discrete phon.</td>
<td>0.646</td>
<td>0.215</td>
<td>0.247</td>
<td>0.377</td>
</tr>
<tr>
<td></td>
<td>discrete dist.</td>
<td>0.572</td>
<td>0.296</td>
<td>0.067</td>
<td>0.309</td>
</tr>
<tr>
<td></td>
<td>continuous dist.</td>
<td>0.694</td>
<td>0.332</td>
<td>0.201</td>
<td>0.465</td>
</tr>
</tbody>
</table>
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!
Discussion

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

- We lose relevant information when we discretize them
<table>
<thead>
<tr>
<th>Model</th>
<th>Featurization</th>
<th>Overall</th>
<th>Attested</th>
<th>Marginal</th>
<th>Unattested</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\tau$</td>
<td>$\tau$</td>
<td>$\tau$</td>
<td>$\tau$</td>
</tr>
<tr>
<td>Smoothed bigram</td>
<td>segments</td>
<td>0.877</td>
<td>0.669</td>
<td>0.509</td>
<td>0.244</td>
</tr>
<tr>
<td>MaxEntGrams</td>
<td>discrete dist.</td>
<td>0.753</td>
<td>0.610</td>
<td>0.424</td>
<td>0.282</td>
</tr>
<tr>
<td>H&amp;W</td>
<td>discrete phon.</td>
<td>0.740</td>
<td>0.674</td>
<td>0.533</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>discrete dist.</td>
<td>0.818</td>
<td>0.634</td>
<td>0.540</td>
<td>0.244</td>
</tr>
<tr>
<td>Bilinear</td>
<td>discrete phon.</td>
<td>0.785</td>
<td>0.646</td>
<td>0.446</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>discrete dist.</td>
<td>0.757</td>
<td>0.572</td>
<td>0.520</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>continuous dist.</td>
<td>0.699</td>
<td><strong>0.694</strong></td>
<td><strong>0.611</strong></td>
<td><strong>0.332</strong></td>
</tr>
</tbody>
</table>

Table 5: Model comparison using Pearson’s $\tau$ and Kendall’s $\tau$ to correlate model scores with acceptability ratings for English onsets. The correlation value for the top performing model in each category is bolded.
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Featurization</th>
<th>Overall</th>
<th>Attested</th>
<th>Marginal</th>
<th>Unattested</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$r$</td>
<td>$\tau$</td>
<td>$r$</td>
<td>$\tau$</td>
</tr>
<tr>
<td>H&amp;W</td>
<td>discrete phon.</td>
<td>0.740</td>
<td>0.674</td>
<td>0.533</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>discrete dist.</td>
<td>0.818</td>
<td>0.634</td>
<td>0.540</td>
<td>0.244</td>
</tr>
<tr>
<td>Bilinear</td>
<td>discrete phon.</td>
<td>0.785</td>
<td>0.646</td>
<td>0.446</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>discrete dist.</td>
<td>0.757</td>
<td>0.572</td>
<td>0.520</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>continuous dist.</td>
<td>0.699</td>
<td><strong>0.694</strong></td>
<td><strong>0.611</strong></td>
<td><strong>0.332</strong></td>
</tr>
</tbody>
</table>

$r =$ Pearson’s rho  \hspace{1cm} $\tau =$ 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

<table>
<thead>
<tr>
<th></th>
<th>sonorant</th>
<th>voice</th>
<th>labial</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>b</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>m</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Put discrete featural representation in a matrix

<table>
<thead>
<tr>
<th></th>
<th>sonorant</th>
<th>voice</th>
<th>labial</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Open question: continuous phonetic feature?

<table>
<thead>
<tr>
<th></th>
<th>sonorant</th>
<th>voice</th>
<th>labial</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Put discrete featural representation in a matrix

<table>
<thead>
<tr>
<th></th>
<th>#_l</th>
<th>#_r</th>
<th>#_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>m</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
PMI

|   | 
|---|---|---|
| p | 2.464 | 1.934 | 0 |
| b | 2.464 | 1.934 | 0 |
| m | 0 | 0 | 0 |