Modeling phonetic category learning from natural acoustic data

Stephanie Antetomaso\textsuperscript{1}, Kouki Miyazawa\textsuperscript{2,4}, Naomi Feldman\textsuperscript{3}, Micha Elsner\textsuperscript{1}, Kasia Hitczenko\textsuperscript{3}, Reiko Mazuka\textsuperscript{4}

\textsuperscript{1}The Ohio State University  \hspace{1cm} \textsuperscript{2}Fairy Devices Inc.
\textsuperscript{3}University of Maryland \hspace{1cm} \textsuperscript{4}RIKEN Brain Science Institute
Input to infants contains variability
Input to infants contains variability

- Yet children acquire phonetic categories of their native language within the first year (Werker & Tees 1984, Polka & Werker 1994)
Input to infants contains variability

- Yet children acquire phonetic categories of their native language within the first year \cite{Werker:1984, Polka:1994}

- Variability is critical for certain types of language learning \cite{Gomez:2002, Rost:2009}
Learning problem
Learning problem

Input:
- acoustics
- lexical information
- etc.
Learning problem

Input:
- acoustics
- lexical information
- etc.

Output:
Knowledge of native phonetic categories
Learning problem

Input:
- acoustics
- lexical information
- etc.

Output:
Knowledge of native phonetic categories
Learning problem

Input:
- acoustics
- lexical information
- etc.

Output:
Knowledge of native phonetic categories

Computational Models
e.g. Vallabha et al. (2007), McMurray et al. (2009), Feldman et al. (2013)
Bayesian lexical-distributional clustering model

Acoustic Values
  Vowel formant pairs

Lexical Context
  Categorical consonant frame

Bayesian Model

Feldman et al. 2013
Bayesian lexical-distributional clustering model

Acoustic Values
- Vowel formant pairs

Lexical Context
- Categorical consonant frame

Vowel Categories
- Gaussian distributions of sound

Lexical Categories
- Sequences of phones

Feldman et al. 2013
Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Bayesian Model

Vowel Categories
Gaussian distributions of sound

Lexical Categories
Sequences of phones

Feldman et al. 2013
Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Feldman et al. 2013

Bayesian Model

Vowel Categories
Gaussian distributions of sound

Lexical Categories
Sequences of phones

ih t s || k a y n d || ah v

F1
430
2100

F2
2100
1300
700
1220

550
1300

700
1220

430
2100

t s || k

550
1300

700
1220

n d || v
Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Bayesian Model

Vowel Categories
Gaussian distributions of sound

Lexical Categories
Sequences of phones

Feldman et al. 2013
Feldman et al. 2013 results

<table>
<thead>
<tr>
<th></th>
<th>Distributional</th>
<th>Lexical-Distributional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.45</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Acoustic simplification: lab productions

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Bayesian Model

Vowel Categories
Gaussian distributions of sound

Lexical Categories
Sequences of phones

Hillenbrand et al. 1995

Feldman et al. 2013
Acoustic simplification: lab productions

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

- Stressed, single syllable
- No noise, reduction, co-articulation
- No **prosodic variability** (affects vowel quality and duration)

Bayesian Model

Vowel Categories
Gaussian distributions of sound

Hillenbrand et al. 1995

Lexical Categories
Sequences of phones

Feldman et al. 2013
Acoustic simplification: corpus vowels

- Acoustic Values
  - Vowel formant pairs

- Bayesian Model

- Vowel Categories
  - Gaussian distributions of sound

- Lexical Context
  - Categorical consonant frame

- Lexical Categories
  - Sequences of phones

Buckeye Speech corpus (Pitt et al. 2007)

Feldman et al. 2013
Lexical simplification: phonemic transcription

- Acoustic Values
  - Vowel formant pairs

- Lexical Context
  - Categorical consonant frame

- Bayesian Model

- Vowel Categories
  - Gaussian distributions of sound

- Lexical Categories
  - Sequences of phones

Feldman et al. 2013

It's kind of a unique position

- ih.t.s k.ay.n.d
- ah.v ey
- y.uw.n.iy.k
- p.ah.z.ih.sh.ah.n
Lexical simplification: phonetic transcription

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Vowel Categories
Gaussian distributions of sound

Bayesian Model

It's kind of a unique position

Lexical Categories
Sequences of phones

Feldman et al. 2013
What Children Actually Hear

Natural Vowels

Reduced Lexical Items

ih.s k.ah.nx
ah.v ey
y.ih.n.iy.k
p.ah.z.ah.sh.ah.n
What Models Actually Receive

Lab Vowels

Phonemic Transcription

ih.t.s k.ay.n.d
ah.v ey
y.uw.n.iy.k
p.ah.z.ih.sh.ah.n
How do models perform given more naturalistic data?

- Models help explore what can be learned from the input, given some algorithm
How do models perform given more naturalistic data?

- Models help explore what can be learned from the input, given some algorithm.
- Computational models aren't people; some simplification to input must be made.
How do models perform given more naturalistic data?

- Models help explore what can be learned from the input, given some algorithm.
- Computational models aren't people; some simplification to input must be made.
- What is the impact of input simplifications on model performance?
How do models perform given more naturalistic data?

• Models help explore what can be learned from the input, given some algorithm

• Computational models aren't people; some simplification to input must be made

• What is the impact of input simplifications on model performance?

• Are conclusions drawn from these models reliable?
Overview

- Simulation 1: Replication of Simplified Input
- Simulation 2: More realistic lexical information
- Simulation 3: More realistic acoustic information
Corpora

• Laboratory vowel productions:
  – English: Hillenbrand et al. 1995
  – Japanese: Mokhtari & Tanaka 2000

• Natural Speech:
  – English: Buckeye Speech corpus (Pitt et al. 2007)
  – Japanese: R-JMICC corpus (Mazuka et al. 2006)
Simulation 1: Replication of Simplified Input

**Acoustic Values**
- Vowel formant pairs

**Lexical Context**
- Categorical consonant frame

**Resampled**

**Bayesian Model**

**Phonemic**
- k.ay.n.d

*Hillenbrand et al. 1995*
*Mokhtari & Tanaka 2000*

*Buckeye Speech corpus (Pitt et al. 2007)*
*R-JMICC corpus (Mazuka et al. 2006)*
Replication of Simplified Input: English vs. Japanese Lab Vowels
Simplified Input:
Successful Category Recovery

English

Japanese
Simplified Input:
Successful Category Recovery

<table>
<thead>
<tr>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories found by model</td>
<td></td>
</tr>
</tbody>
</table>

Categories:**aa, ae, ah, eh, er, ey, ih, iy, ow, uh, uw**

True Categories:**/i/, /e/, /a/, /o/, /u/**
Simplified Input: Successful Category Recovery
Simplified Input:
Successful Category Recovery

Phonetic F-Score: 0.78

Original Feldman et al. 2013 results: Phonetic F-Score: 0.76
Simplified Input: Successful Category Recovery

Phonetic F-Score: 0.78

Phonetic F-Score: 0.98

Original Feldman et al. 2013 results: Phonetic F-Score: 0.76
Simulation 2: Phonetic Transcription

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Resampled

Bayesian Model

Phonetic
k.ah.nx
Corpus effects of phonetic transcription

- English: vowels of frequent words reduced to schwa in natural speech $\rightarrow$ increased number of phonetic variants

\[
\begin{align*}
\text{k.ae.n} & \rightarrow \text{k.ae.n} \\
\text{k.ih.n} & \rightarrow \text{k.ih.n} \\
\text{k.eh.n} & \rightarrow \text{k.eh.n} \\
\text{k.ah.n} & \rightarrow \text{k.ah.n} \\
\text{k.n} & \rightarrow \text{k.n} \\
e tc. & \rightarrow \text{etc.}
\end{align*}
\]
Corpus effects of phonetic transcription

- English: vowels of frequent words reduced to schwa in natural speech $\rightarrow$ increased number of phonetic variants
- Japanese: less phonetic reduction

<table>
<thead>
<tr>
<th>English Word Types</th>
<th>Japanese Word Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonemic Transcription: 1099</td>
<td>Phonemic Transcription: 751</td>
</tr>
<tr>
<td>Phonetic Transcription: 1813</td>
<td>Phonetic Transcription: 791</td>
</tr>
</tbody>
</table>
Phonetic Transcription:
Decline in English Performance
Phonetic Transcription: Decline in English Performance

Phonetic F-Score: 0.46

Phonetic F-Score: 0.95
Phonetic Transcription:
Decline in English Performance

Phonetic F-Score: 0.46

Phonetic F-Score: 0.95
Simulation 3: Realistic Vowels From Corpus

- Acoustic Values: Vowel formant pairs
- Lexical Context: Categorical consonant frame
- Corpus Vowels
- Phonetic: k.ah.nx
- Bayesian Model
Simulation 3: Realistic Vowels From Corpus

English

Japanese
Realistic Vowels:
Poor Category Recovery

English

Japanese
Realistic Vowels: Poor Category Recovery

English

Japanese

Phonetic F-Score: 0.13

Phonetic F-Score: 0.22
Realistic Vowels: Poor Category Recovery

Phonetic F-Score: 0.13

Phonetic F-Score: 0.22
Japanese phrase-final lengthening

• Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama

mama:
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama

mama:
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama

mama:
## Summary of Results

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Phonetic F-Score English</th>
<th>Phonetic F-Score Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 1: Simplified Input</td>
<td>0.78</td>
<td>0.98</td>
</tr>
<tr>
<td>Simulation 2: Phonetic Transcription</td>
<td>0.46</td>
<td>0.95</td>
</tr>
<tr>
<td>Simulation 3: Realistic Vowels</td>
<td>0.13</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Discussion

• There is little variability in simplified input, but a lot in the input received by children
  – Lexical variability
  – Acoustic variability

• Adding this variability back to the input can drastically impact model performance, and may have different effects on different languages.

• To explore the learning problem we must have ecologically valid datasets
Thank You!

NSF IIS-1422987
NSF IIS-1421695
NSF DGE-1343012
OSU Lacqueys reading group
Japanese Long vs Short

Figure 1. Mean duration of short and long vowels in the present Japanese IDS corpus. The difference in duration between short and long vowels is reliable and the effect size is large. The error bars represent the standard error of the mean for each vowel across participants. doi:10.1371/journal.pone.0051594.g001
English versus Japanese Duration

Gaussian mixture models (e.g., Vallabha, McClelland, Pons, Werker, & Amano, 2007; McMurray, Aslin, & Toscano, 2009)
CDS versus ADS
Speaker Variability
## Summary of results

<table>
<thead>
<tr>
<th></th>
<th>Phonetic F-score</th>
<th>Lexical F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Japanese</td>
</tr>
<tr>
<td></td>
<td>AD</td>
<td>CD</td>
</tr>
<tr>
<td>Corpus 1</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>Corpus 2</td>
<td>0.46</td>
<td>-</td>
</tr>
<tr>
<td>Corpus 3</td>
<td>0.13</td>
<td>-</td>
</tr>
</tbody>
</table>
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama:
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama: