Learning a variable language

Micha Elsner
Department of Linguistics

THE OHIO STATE UNIVERSITY
Learning a language...

http://childes.psy.cmu.edu/media/Eng-NA-MOR/Providence/Lily/lil10.mp4

“Where’s Birthday Bear?”

The infant must learn:

- \textit{bear} is a word that means
- \textit{bear} and \textit{where} are different words…
- English /\textit{b}/ and /\textit{w}/ are different sounds
Modeling: learn by predicting

What do you need to predict a missing item?

Good hypothesized words make ‘where’ predictable

Bad hypothesized words make the gap hard to fill in

Bigram Markov model:
Goldwater, Griffiths and Johnson 2009
Assumption: variability is signal

Different contexts mark different words
Bayesian model of learning

\[ P(h|D) \propto P(h)P(D|h) \]

support for hypothesis \( h \) given dataset \( D \)

prior on \( h \):
preference for small number of relatively simple categories
(not too many words)

likelihood of \( D \):
preference for predictive power about the data
(fill in gaps easily)

\( h \) encodes any robustly predictable variability in \( D \)
Problem: variability is all over

One speaker’s pronunciations differ:

- Speech rate
- Pitch
- Lexical context

And among speakers:

- Gender
- Dialect
- And much more!
Models try to encode *all* this variability

Different speech styles can be interpreted as different words
Overlearning in a real model

Vertical lines on figure:
real sound category distributed among multiple model categories

Vowels and sibilants are split among multiple classes:
Contextually dependent variants treated as categories
Explicitly model variability

“about”

about  erbout  bout

at about  make about  ...

...
Overview:

- **Word learning with variable pronunciations**
  - Grouping *you* and *ya*
- **Learning words and acoustics together**
  - Learning */u/* and *you*
- **Steps toward naturalistic datasets**
  - Model degradation on conversational speech
  - Forced alignment: getting more data cheaply
Word forms with variability
Elsner, Goldwater, Feldman and Wood 2013
Context: word segmentation

you want to see the book look there’s a boy with his hat and a doggie
you want to look at this look at this have a drink take it out you want it in put that on that
yes okay open it up take the doggie out i think it will come out what daddy where did it go you want that one daddy i’ll go get your block what’s that alice what’s that block that’s a telephone that’s the phone say hello
you want to speak to alice say hello what’s you have to tell me block
you want the blocks

- The infant hears a stream of utterances
- And has to pick out repeated units
Experimental evidence

Stress and intonation play a role (Thiessen and Saffran ‘03)
But mere repetition can be enough (Saffran et al ‘96)
Especially near a familiar ‘anchor’ word (Bortfeld et al ‘05)

“Where’s [Birthday] Bear?”
Classic model setup (Brent '99)

**Human transcriber**
(audio : sentence)

**Deterministic dictionary**
(word : phonemes)

you \(\text{ju}\) want want ...

**Deterministic dictionary**
(phonemes : word)

you \(\text{ju}\) want want ...

**Language model**
(words : sentence)

you want you like ...

---

**Audio**

**Orthographic**

you want a cookie

**Normalized phonetic transcript**

\(\text{juwantækuki}\)

**Segmented**

you want a cookie

**Experimental setup**

**Model inference**
This setup masks variability

- Words given unrealistic ‘citation form’ pronunciations
- Token frequencies inflated
  - By grouping together tokens that sound different
- Model doesn’t need to learn any phonology
Variable input in real life

In word learning simulation, infants (19 months):

- accept near variants of new word
- but not more distinct ones

White and Morgan 2008
Developmental pathway

Results suggest:

● Initial problems adapting to tone of voice or speaker identity (Houston and Jusczyk, 2000; Singh et al., 2004)
● Vowel contrasts emerge earlier (Kuhl et al., 1992; Bosch and Sebastian-Galles, 2003; Werker and Tees, 1984)
● Words in middle of phrase are harder (Plunkett, 2005; Seidl and Johnson, 2006)
  ○ Suggesting segmentation issues?
Our setup

Experimental setup

Model inference

Human transcriber (audio : sentence)

Pronunciation dictionary (word : phonemes)

you \quad ju, jə, ji
want \quad wan, want
...
Variants (from Buckeye corpus)

“about”

ahbawt: 15, bawt: 9, ihbawt: 4, ahbawd: 4, ihbawd: 4, ahbaat: 2, baw: 1, ahbaht: 1, erbawd: 1, bawd: 1, ahbaad: 1, ahpaat: 1, bah: 1, baht: 1

“gonna”

Noisy channel model

- Probabilistically generate canonical words
- Then rewrite them into observed forms

random lexicon
- want, ju...
- word-to-word transition probabilities
  \[ p(\text{want}|\text{ju}), \ p(\text{to}|\text{want}) \]

intended utterances
- ju want wan
- want e kɔki

noisy channel
- character sequence rewrite probabilities
  \[ p(u \rightarrow a : j_\$) \]

surface (observed)
- ja wa? wəŋ
- wan a ə kɔki
Dense model description

finite-state transducer: independent single-segment substitutions

Generator for possible words
a, b, ..., ju, ... want, ... ju want, ...

Probabilities for each word
(sparse)
p(δi) = .1, p(a) = .05, p(want) = .01...

Conditional probabilities for each word after each word
p(δi | want) = .3, p(a | want) = .1,
p(want | want) = .0001...

Intended forms
ju want a kuki
ju want it
...

Surface forms
ja wan a kuki
ju wand it
...
Inference: finite-state encoding

(Not quite exact; can be corrected with Metropolis-Hastings)

(van Gael et al ‘08; Huggins and Wood ‘13)
Beam sampling

Starting with current trajectory...

\[
\begin{array}{c}
\text{[s]} \quad \overset{j/j}{\longrightarrow} \quad \overset{u/u}{\longrightarrow} \quad j \quad \overset{d}{\longrightarrow} \quad \overset{j/k}{\longrightarrow} \quad k \quad \overset{\theta}{\longrightarrow} \quad \text{word } j\theta
\end{array}
\]

\[p(j\theta|[s])\]

(van Gael et al ‘08; Huggins and Wood ‘13)
Beam sampling

Sample stochastic cutoff points to limit search space

(van Gael et al ‘08; Huggins and Wood ‘13)
With variation, fewer bogus “words”

Words containing “you” from our model:
  you (805 times), doyou (240 times), youwan (88 times), yih (58 times), areyou (54 times), youdo (47 times)

Words containing “you”; no phonetic variation:
  you (498 times), yih (280 times), ya (165 times), yee (119 times),
  doyou (106 times), doyee (44 times), canyou (39 times), canyee (29 times)

Our model learns a compact early lexicon
  • Like real infants, accepts variants of words
## Learned rules for variation

<table>
<thead>
<tr>
<th>System</th>
<th>Intended</th>
<th>Top outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>u</td>
<td>ŭ .68</td>
</tr>
<tr>
<td></td>
<td>ð</td>
<td>ð .69</td>
</tr>
<tr>
<td>Learned</td>
<td>u</td>
<td>ŭ .75</td>
</tr>
<tr>
<td></td>
<td>ð</td>
<td>ð .91</td>
</tr>
</tbody>
</table>

- System learns plausible sound changes
- But overconservative ones…
- Model of vowels is closer to reality
# Vowel-initial words are hard

<table>
<thead>
<tr>
<th></th>
<th>Vowel-initial</th>
<th>Cons-initial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>41.5</td>
<td>52.1</td>
</tr>
<tr>
<td>Segmented correctly, but wrong word: in / an</td>
<td>20.4</td>
<td>17.3</td>
</tr>
<tr>
<td>Collocation: Missed boundary with neighboring word: that’s a / that’s</td>
<td>19.2</td>
<td>12.5</td>
</tr>
</tbody>
</table>
These results suggest...

- Infants’ early conservatism about sound shifts may keep word learning tractable
- Vowels are easier because lexical evidence is better
- Vowel-initial words prone to both misrecognition and mis-segmentation
Weaknesses

Although the data here is more naturalistic:

- Discrete symbols, not acoustics
- Only segment-by-segment variation
- *Plausible* variations in data, not directly transcribed
  - Due to expense of closely transcribing child-directed speech
Words and phonetic categories

Elsner, Feldman and Antetomaso, work in progress
Feldman, Griffiths and Morgan, 2009
A step towards acoustics

Replace vowel symbols with continuous measurements:

Before:
\[ y\ uw \ || \ w\ aa\ n \ || \ t\ uw \ || \ s\ iy \ || \ dh\ iy \ || \ b\ uh\ k \ || \]

Now:
\[ y <380.5\ 1251.6> \ || \ w <811.8\ 1431.9>\ n \ || \ t <532.9\ 1094.1> \ || \]
\[ s <468.2\ 2703.2> \ || \ dh <595.2\ 973.8> \ || \ b <545.3\ 1330.0>\ k \ || \]
Vowels characterized by *formants* (resonances of the vocal tract)

- Can be plotted and clustered in two dimensions
Formant space isn’t enough

Vowel categories overlap

A: laboratory vowel data (Hillenbrand et al ‘95)
B: max-likelihood mixture of Gaussians (Vallabha et al ‘07)
C: Bayesian infinite GMM (Feldman et al ‘09)
Feldman: helps to learn lexicon

Vowel data

Learned by lexical model

Can also benefit from semantics (Frank et al ‘14)

But can we do it jointly with word segmentation?
Joint segmentation model

Not that hard to specify:
• Replace **discrete transducer** with **infinite mixture of Gaussians**

Inference is more complex:
• Feldman sampler relies on block moves
• Harder to implement in bigram model

*Caveat for these runs: fixed # components*
Joint model results

Word segmentation scores decline a bit:

<table>
<thead>
<tr>
<th>Unlabeled surface token score</th>
<th>Token P</th>
<th>Token R</th>
<th>Token F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goldwater (discrete vowels)</td>
<td>75</td>
<td>72</td>
<td>73</td>
</tr>
<tr>
<td>Joint model</td>
<td>66</td>
<td>71</td>
<td>68</td>
</tr>
</tbody>
</table>

Category learning scores too:

<table>
<thead>
<tr>
<th>Cluster micro-F</th>
<th>Vowel P</th>
<th>Vowel R</th>
<th>Vowel F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feldman-like (no segmentation)</td>
<td>84</td>
<td>82</td>
<td>83</td>
</tr>
<tr>
<td>Joint model</td>
<td>91</td>
<td>77</td>
<td>83</td>
</tr>
</tbody>
</table>
Convergence over time

System clearly identifies vowels, even poorly-separated ones
Weaknesses

● Joint sampler mixes poorly
  ○ At low temperature, hard to explore new vowels
  ○ At high temperature, too many vowels limits speed
● *Still* not natural data!
Towards more realistic corpora
Antetomaso and Elsner, in progress
Austen and Elsner, in submission
Let’s finally get some real data

Recorded vowels from the Buckeye corpus

(Sadly, not child-directed)

(Pitt et al ‘07)
Uncontrolled variability!

- Between speakers
  - Including gender and age
- Coarticulation
  - Contexts include nasals, liquids, etc.
- Reduction and other prosodic effects

How bad is this for the model?
But what exactly will go wrong?

A series of experiments:

- Running the Feldman ‘09 vowel learner
  - No segmentation
- With different lexicons
- And different acoustic values
<table>
<thead>
<tr>
<th>Unigram frequencies</th>
<th>Pronunciations</th>
<th>Acoustic coordinates</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a: child-directed</td>
<td>canonical pronunciation</td>
<td>resampled from Hillenbrand</td>
<td>“kind” : k ay n d</td>
</tr>
<tr>
<td>b: Buckeye</td>
<td>transcribed (variable) pronunciation</td>
<td>resampled from Hillenbrand</td>
<td>“kind” : k ah nx</td>
</tr>
<tr>
<td>Buckeye</td>
<td>transcribed (variable) pronunciation</td>
<td>actual acoustics measured <em>in situ</em></td>
<td>“kind” : k ah nx</td>
</tr>
</tbody>
</table>
**Buckeye lexicon works fine**

<table>
<thead>
<tr>
<th>a: child-directed</th>
<th>canonical pronunciation</th>
<th>resampled from Hillenbrand</th>
<th>“kind” : k ay n d</th>
</tr>
</thead>
<tbody>
<tr>
<td>b: Buckeye</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lexicon F: a: 93 b: 96 (1018 words, gold: 961)

<table>
<thead>
<tr>
<th>Buckeye</th>
<th>transcribed (variable) pronunciation</th>
<th>resampled from Hillenbrand</th>
<th>“kind” : k ah nx</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(variable)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Buckeye</th>
<th>transcribed (variable) pronunciation</th>
<th>actual acoustics measured in situ</th>
<th>“kind” : k ah nx</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Transcribed variation breaks stuff

<table>
<thead>
<tr>
<th>Lexicon F: a: 93 b: 96 (1018 words, gold: 961)</th>
<th>Phone F: a: 76, b: 78</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buckeye transcribed (variable) pronunciation</td>
<td>resampled from Hillenbrand</td>
</tr>
<tr>
<td>Lexicon F: 63 (1355 words, gold: 1813)</td>
<td>Phone F: 46</td>
</tr>
<tr>
<td>Buckeye transcribed (variable) pronunciation</td>
<td>actual acoustics measured <em>in situ</em></td>
</tr>
</tbody>
</table>

- **Child-directed** pronunciation: 
  - *Buckeye* canonical pronunciation resampled from Hillenbrand: “kind” : k ay n d
  - *Buckeye* transcribed (variable) pronunciation resampled from Hillenbrand: “kind” : k ah nx
  - *Buckeye* transcribed (variable) pronunciation actual acoustics measured *in situ*: “kind” : k ah nx
What goes wrong?

Pronunciation variation is a big deal:

- Increases number of lexical items
- Increases number of different vowels occurring in same context
Lab vowels with variable lexicon

- “Wastebasket” categories capture multiple variants of words like “you” and “the”
- Severely underestimates lexicon size
## Real phonetics are a disaster

<table>
<thead>
<tr>
<th>a: child-directed</th>
<th>canonical pronunciation</th>
<th>resampled from Hillenbrand</th>
<th>“kind” : k ay n d</th>
</tr>
</thead>
<tbody>
<tr>
<td>b: Buckeye</td>
<td></td>
<td></td>
<td>“kind” : k ah nx</td>
</tr>
<tr>
<td>Lexicon F: a: 93</td>
<td></td>
<td></td>
<td>“kind” : k ah nx</td>
</tr>
<tr>
<td>b: 96 (1018 words, gold: 961)</td>
<td></td>
<td></td>
<td>“kind” : k ah nx</td>
</tr>
<tr>
<td>Phone F: a: 76,</td>
<td></td>
<td></td>
<td>“kind” : k ah nx</td>
</tr>
<tr>
<td>b: 78</td>
<td></td>
<td></td>
<td>“kind” : k ah nx</td>
</tr>
</tbody>
</table>

Buckeye transcribed (variable) pronunciation resampled from Hillenbrand “kind” : k ah nx

Lexicon F: 63 (1355 words, gold: 1813) Phone F: 46

Buckeye transcribed (variable) pronunciation actual acoustics measured in situ “kind” : k ah nx

Lexicon F: 41 (1381 words, gold: 1882) Phone F: 13
Fit to in-situ vowels

- 20 inferred vowels
- High degree of overlap
Conclusion

Model needs both:
- Discrete (phonological) variation
- Continuous (phonetic) variation

(cf Lee 2014)

The difference between “you” and “yih”,
Versus high /u/ and low /u/
Getting better data

Buckeye is *adult*-directed speech

Two-pronged effort:

- Automatically time-align child-directed speech
- Use existing Japanese resource \(\text{(Mazuka et al '06)}\)
Better forced alignment tools

- We have plenty of transcribed child-directed speech
- But it’s hard to use for cognitive models
- We don’t know what sounds are where!

Forced alignment: speech recognizer finds vowels by lining up with the transcript
Except...

Forced alignment often goes wrong
Incorrect segment detection

<table>
<thead>
<tr>
<th></th>
<th>Error detection rate</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>System</td>
<td>61</td>
<td>34</td>
</tr>
</tbody>
</table>

- CRF with hand-tuned and neural net acoustic features
- 77% of errors for chains of 11 or more
Overall conclusions

- Using lab data in learning simulations can mask real and interesting variation
- Learning models should model this variation

Child errors may be due to difficult balance between learning sound categories and words
Thanks to...

NSF 1422987 “Cognitive models of the acquisition of vowels in context”, with Naomi Feldman

OSU Lacqueys reading group
Mary Beckman, Marie-Catherine de Marneffe and Laura Wagner

And all of you!