Sounds to Words
Bridging the Gap

Micha Elsner
with Sharon Goldwater,
Jacob Eisenstein and Frank Wood

Department of Linguistics
The Ohio State University
University of Edinburgh, Georgia Tech
and Columbia University

October 30, 2012
Early language learning

Audio

Interpretable

Early language learning

Phonetic categories (audio : phoneme)

Pronunciation dictionary (phonemes : word)

Language model (words : sentence)

Audio

Phonetic transcript

Interpretable

Segmented

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/

you

you want

you like

...
Are /u/ and /i/ different vowels?

- Pronunciation: yes, because /ju/ is common and /ji/ is rare
Are /u/ and /i/ different vowels?
  - Pronunciation: yes, because /ju/ is common and /ji/ is rare

Is /ju/ a word?
  - Phonetics: yes, because /j/ predicts /u/-like vowel tokens
  - Language model: yes, because it helps predict subsequent /want/
Are /u/ and /i/ different vowels?
- Pronunciation: yes, because /ju/ is common and /ji/ is rare

Is /ju/ a word?
- Phonetics: yes, because /j/ predicts /u/-like vowel tokens
- Language model: yes, because it helps predict subsequent /want/

Is /ju/ /want/ different from /ðɛj/ /want/?
- Pronunciation: yes, because they contain dissimilar segments

Components interact to solve the problem...
Evidence from development

- **Phonetics**
  - Native vowel contrasts (Polka+Werker 94)
  - Native consonant contrasts (Werker+Tees 84)
  - Frequent words (Bergelson+Swingley 12)
  - Frequent words (Jusczyk+al 95, 99)
  - Names (Bortfeld+al 05)
  - Function words (Shady 96)

- **Lexicon**
  - Birth
  - 6 months
  - 8 months
  - 1 year

Key developments at roughly the same time following presentations by Feldman 09, Dupoux 09
Want to show: these synergies are real!

Cognitive/Linguistic

- Establish role of synergy in early acquisition
- Propose mechanisms: predict developmental stages
- Universals vs. generic learning
Want to show: these synergies are real!

Cognitive/Linguistic
- Establish role of synergy in early acquisition
- Propose mechanisms: predict developmental stages
- Universals vs. generic learning

Applied
- Unsupervised speech recognition
- Learn new lexical items/accents
Related work

- (Martin, Peperkamp, Dupoux ‘12) Clusters symbolic phones into phonemes by learning a proto-lexicon
- (Feldman, Griffiths, Morgan ‘09) Clusters acoustic tokens into phonemes based on a known lexicon
- (Plaut, Kello ‘98) Neural network model of phonetic articulations from known lexicon, uncertain semantics
- (Neubig et al ‘12), (Rytting, Brew 2008) Learn words given uncertain representation of input
- (Vallabha+al ‘07, Varadarajan+al ‘08, Dupoux+al ‘11, Lee+Glass ‘12) Discover phone-like units from acoustics (no lexicon)
Related work

- (Martin, Peperkamp, Dupoux ‘12) Clusters symbolic phones into phonemes by learning a proto-lexicon
- (Feldman, Griffiths, Morgan ‘09) Clusters acoustic tokens into phonemes based on a known lexicon
- (Plaut, Kello ‘98) Neural network model of phonetic articulations from known lexicon, uncertain semantics
- (Neubig et al ‘12), (Rytting, Brew 2008) Learn words given uncertain representation of input
- (Vallabha+al ‘07, Varadarajan+al ‘08, Dupoux+al ‘11, Lee+Glass ‘12) Discover phone-like units from acoustics (no lexicon)

This work

Large, semi-realistic corpus of symbolic input
Learns explicit lexicon and phonetic rules
Future work could integrate some other models!
In this talk

Motivation

Word segmentation: previous work on the lexicon
   Goldwater’s Bayesian model of lexical acquisition

Modeling phonetic variation (ACL ‘12)
   Our Bayesian model
   Channel model: transducer with articulatory features
   Bootstrapping the model
   Greedy inference
   Performance

Jointly segmenting and modeling variation
   Inference with beam sampling

Conclusions
Word segmentation
(Setup follows (Brent ‘99))

Human transcriber
(audio : sentence)

Deterministic dictionary
(word : phonemes)

you  ju
want  want ...

Deterministic dictionary
(phonemes : word)

you  ju
want  want ...

Language model
(words : sentence)

you want you like ...

Audio

Orthographic

you want a cookie

Normalized phonetic transcript

ju wantəkuki

Segmented

ju want instead of jə wan

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/
The input

Input from phonetic dictionary: why?

- Pipeline model?
  - Learn phonetics first
  - Use learned phonetics to normalize input

- Little theoretical justification for this...

- Real phonetic transcription is *expensive*!
  - Usually requires linguists
  - Very time-consuming
  - Some for adult speech, no child-directed corpora to my knowledge

Mostly a matter of convenience!
Segmenting words: previous work

Previous models use two kinds of evidence:

**Boundary-based**

/pɛtkɪti/: *tk so /pɛt/ /kɪti/

- Learn about phonotactics
- Place boundaries to break infrequent sound sequences
- Words defined implicitly by boundary position

(Fleck ‘08, Rytting ‘07, Daland+al ‘10, others)

**Lexical**

/pɛtkɪti/: *kɪt* probably a word, so /pɛt/ /kɪti/

- Learn probable lexical items
- Propose word sequence to cover observed corpus
- Boundaries defined implicitly by word sequence

(Brent ‘99, Venkataraman ‘01, Goldwater ‘09, others)
Goldwater et al ‘09

A lexical model of word segmentation:

- Generative Bayesian model
- Two parts: probability of lexicon
  - Dirichlet process: allows infinite, favors small
- Probability of corpus
  - Rewards predictability
- Basis for other work in this talk
Generative story

Infinite list of possible words
  a, b, ..., ju, ... want, ... juwant, ...

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

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

Observed corpus
  ju want ə kuki
  ju want ɪt
  ...

α
θ
0

α
θ

∞ contexts

n utterances x1 x2 ...
What’s going on?

**Memorizing the data**
Lexicon: *juwantəkuki, juwantit*
Likelihood of corpus is high...
But lexicon is huge: sparse prior says not very likely

**Character by character**
Lexicon: *j, u, w...*
Lexicon is very sparse: prior is high
Likelihood of corpus is poor

**True lexicon**
Lexicon: *ju, want...*
A “happy medium”
Goldwater suffers under variation

Goldwater run on Buckeye corpus (Fleck ‘08)
- Must represent each pronunciation separately
- No var. (dictionary) versus phonetic transcript

<table>
<thead>
<tr>
<th>Break F</th>
<th>Token F</th>
<th>Lexicon (type) F</th>
</tr>
</thead>
<tbody>
<tr>
<td>no var.</td>
<td>84</td>
<td>68</td>
</tr>
<tr>
<td>transcribed</td>
<td>65</td>
<td>35</td>
</tr>
</tbody>
</table>

Break F declines: huge decrease in precision
- Undersegmentation
Overview

Motivation

Word segmentation: previous work on the lexicon
  Goldwater’s Bayesian model of lexical acquisition

Modeling phonetic variation (ACL ‘12)
  Our Bayesian model
  Channel model: transducer with articulatory features
  Bootstrapping the model
  Greedy inference
  Performance

Jointly segmenting and modeling variation
  Inference with beam sampling

Conclusions
(Simple) phonetic variation

**Human transcriber** (audio : sentence)

**Pronunciation dictionary** (word : phonemes)

**Pronunciation dictionary** (phonemes : word)

**Language model** (words : sentence)

Audio

Orthographic

Segmented phonetics

Segmented

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/
Noisy channel setup

- **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 kōki*

- **noisy channel**
  - character sequence rewrite probabilities
    - $p(u \rightarrow \varepsilon : j_\$)$

- **surface (observed)**
  - *ja wa? wan*
  - *wan e kōki*
Graphical model

Presented as Bayesian model to emphasize similarities with (Goldwater+al ‘09)
  - Our inference method approximate
Graphical model

\[ \alpha \rightarrow \theta \rightarrow \pi(\delta_i) \rightarrow \pi(\text{want}) \rightarrow \ldots \]

Dirichlet process

\[ \alpha \rightarrow X \rightarrow \delta_i \rightarrow \theta \rightarrow \pi(\delta_i) \rightarrow \pi(\text{want}) \rightarrow \ldots \]

\[ \delta_i \rightarrow \pi(\delta_i) \rightarrow \pi(\text{want}) \rightarrow \ldots \]

\[ \text{S} \rightarrow \text{T} \]

\[ \text{di} \rightarrow \delta_i \rightarrow \theta \rightarrow \pi(\delta_i) \rightarrow \pi(\text{want}) \rightarrow \ldots \]
Graphical model

\[ \alpha \rightarrow \theta \]

\[ l \xleftarrow{\text{want}} \; \xrightarrow{\text{di}} \; s \]

\[ x \xrightarrow{\text{di}} \; \xleftarrow{\text{\ddot{o}i}} \; r \xrightarrow{\text{kuki}} \]

\[ T \]
Graphical model

Pitman-Yor process

\[ p(\text{k\textsc{uki}} \mid \delta_i) \]
\[ p(\text{b\textsc{ol}} \mid \delta_i) \]
\[ p(\text{d\textsc{ogi}} \mid \delta_i) \]
\[ \ldots \]
Minor point: here we factor:

\[ p(l, x, y) = p(x)p(l|x)p(r|x) \]

This generates words twice if we look at the whole corpus...
In this section we only look at subsets of words. Later we switch to:

\[ p(l, x, y) = p(l)p(x|l)p(r|x) \]
Weighted Finite-State Transducer

Reads an input string
Stochastically produces an output string
Distribution $p(out|in)$ is a hidden Markov model

Identity FST given $\delta i$
(reads $\delta i$ "the" and writes $\delta i$)
Our transducer

Produces any output given its input
Allows insertions/deletions

Reads $\odot i$, writes anything
(Likely outputs depend on parameters)
Probability of an arc

How probable is an arc? $[\cdot \delta \ i ] \xrightarrow{\delta/d}$

**Log-linear model**

Extract features $f$ from state/arc pair...

▶ Score of arc $\propto \exp(w \cdot f)$ following (Dreyer+Eisner '08)

**Articulatory features**

▶ Represent sounds by how produced

▶ Similar sounds, similar features
  ▶ $\delta$: voiced dental fricative
  ▶ $d$: voiced alveolar stop

see comp. optimality theory systems (Hayes+Wilson ‘08)
Feature templates for state (prev, curr, next) → output

Templates for voice, place and manner

Ex. template instantiations:

(fric)→stop
(ð)→stop
(voiced)→voiced
(ð)→voiced
same-voicing
(dental)→alveol.
(ð)→alveol.

(fric,vowel)→stop
(ð,vowel)→stop
...

→d
Learned probabilities

\[ \delta \rightarrow i \]

- $\delta = 0.7$
- $n = 0.13$
- $\theta = 0.04$
- $d = 0.02$
- $z = 0.02$
- $s = 0.01$
- $\epsilon = 0.01$

... ...
Inference

Bootstrapping

Initialize: surface type $\rightarrow$ itself ($[\text{di}] \rightarrow [\text{di}]$)

Alternate:

- Greedily merge pairs of word types
  - ex. intended form for all $[\text{di}] \rightarrow [\ddi]$
- Reestimate transducer
Inference

Bootstrapping

Initialize: surface type $\rightarrow$ itself ($[\text{di}] \rightarrow [\text{di}]$)

Alternate:
- Greedily merge pairs of word types
  - ex. intended form for all $[\text{di}] \rightarrow [\text{ði}]$
- Reestimate transducer

Greedy merging step

Relies on a score $\Delta$ for each pair:
- $\Delta(u, v)$: approximate change in model posterior probability from merging $u \rightarrow v$
- Merge pairs in approximate order of $\Delta$
 Computing $\Delta$

$\Delta(u, v)$: approximate change in model posterior probability from merging $u \rightarrow v$

- **Terms from language model**
  - Encourage merging frequent words
  - Discourage merging if contexts differ
  - See the paper

- **Terms from transducer**
  - Compute with standard algorithms
  - (Dynamic programming)

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

intended utterances
ju want wan
want e koji

noisy channel
character sequence rewrite probabilities
$p(u \rightarrow a : j, \$)$

surface (observed)
ja wa? wan
wan e koji
Dataset

We want: child-directed speech, close phonetic transcription

Use: Bernstein-Ratner (child-directed) (Bernstein-Ratner ’87)

Buckeye (closely transcribed) (Pitt+al ‘07)

Sample pronunciation for each BR word from Buckeye:

- No coarticulation between words

“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
Evaluation

Map system’s proposed intended forms to truth

- \{\text{ði, di, ðə}\} cluster can be identified by any of these
- System doesn’t do “phonology”— at this stage, neither may infant?
- Score by tokens; emphasis on frequent words
- ...and types (lexicon); all lexemes counted equally
With gold segment boundaries

### Scores (correct forms)

<table>
<thead>
<tr>
<th></th>
<th>Token F</th>
<th>Lexicon (Type) F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ident)</td>
<td>65</td>
<td>67</td>
</tr>
<tr>
<td>Initializer</td>
<td>75</td>
<td>78</td>
</tr>
<tr>
<td>No context</td>
<td>75</td>
<td>76</td>
</tr>
<tr>
<td><strong>Full system</strong></td>
<td><strong>79</strong></td>
<td><strong>87</strong></td>
</tr>
<tr>
<td>Upper bound</td>
<td>91</td>
<td>97</td>
</tr>
</tbody>
</table>
Learning

Initialized with weights on *same-sound, same-voice, same-place, same-manner*

![Graph showing learning progress with iterations and token lexicon metrics.](image-url)
Induced word boundaries

Induce word boundaries with (Goldwater+al ‘09)
Cluster with our system

Scores (correct boundaries and forms)

<table>
<thead>
<tr>
<th></th>
<th>Token F</th>
<th>Lexicon (Type) F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ident)</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>Full system</td>
<td>49</td>
<td>46</td>
</tr>
</tbody>
</table>

After clustering, remove boundaries and resegment: no improvement
Suggests joint segmentation/clustering
Overview

Motivation

Word segmentation: previous work on the lexicon
   Goldwater’s Bayesian model of lexical acquisition

Modeling phonetic variation (ACL ‘12)
   Our Bayesian model
   Channel model: transducer with articulatory features
   Bootstrapping the model
   Greedy inference

Performance

Jointly segmenting and modeling variation
   Inference with beam sampling

Conclusions
Joint segmentation and word forms

- **Human transcriber** (audio : sentence)
- **Pronunciation dictionary** (word : phonemes)
  - you, ju, jə, jɪ
  - want, wan, want
  - ...
- **Pronunciation dictionary** (phonemes : word)
  - jəwanəkuki
- **Language model** (words : sentence)
  - you want
  - you like
  - ...

Audio

**Orthographic**

**Phonetic transcript**

**Segmented**

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/
Challenges

Model from previous section fine for joint segmentation/clustering
  ▶ (Factor $p(l)p(x|l)p(r|x)$ but this is trivial fix)

Issue is inference:
Challenges

Model from previous section fine for joint segmentation/clustering

- (Factor $p(l)p(x|l)p(r|x)$ but this is trivial fix)

Issue is inference:

- Standard segmentation: sample locations of boundaries
- Only two steps to slice out *want*
- *juwanttu*, *ju•wanttu*, *ju•want•tu*
Challenges

Model from previous section fine for joint segmentation/clustering

- (Factor $p(l)p(x|l)p(r|x)$ but this is trivial fix)

Issue is inference:

- Standard segmentation: sample locations of boundaries
- Only two steps to slice out want
- $j\dot{u}\dot{w}anttu$, $j\ddot{u}\ddot{w}anttu$, $j\dddot{u}\dddot{w}ant\dddot{u}$

- With clustering, have farther to travel
- $j\dot{w}antu$, $j\ddot{w}antu$, $j\dddot{w}an\dddot{t}u$, $j\dddot{u}\dddot{w}an\dddot{t}u$, $j\dddot{u}\dddot{w}ant\dddot{t}u$
- Need moves that alter multiple letters/boundaries at once
Markov-style sampling methods

Can write a Goldwater model as a big FSM:

(First in (Mochihashi+al '09))

Only need states that generate the original string

Use forward-backward (plus MH) for inference
Composing with transducer

Unigram transducer
(read any, write any)

Acceptor for original string
Result is a REALLY BIG transducer

[ s ] → j → u → word j

j/j

p(j|[s])

u/u

p(u|j)

word u

j u

word ju

p(ju|[s])

word ju

j/đ

j/d

d

...
Sampling from huge transducers (beam sampling)

\[ \text{[s]} \rightarrow j \rightarrow u \]

\[ j/j \quad u/u \]

\[ \text{word } j\text{\textsc{\ae}} \]

\[ \text{p}(j\text{\textsc{\ae}}|[s]) \]

(van Gael+al ‘08), (Huggins+Wood ‘12)
Sampling from huge transducers (beam sampling)

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

\[ \text{word } j\theta \]

(van Gael+al ‘08), (Huggins+Wood ‘12)
Sampling from huge transducers (beam sampling)

\( [s] \rightarrow j \rightarrow u \)

\( \sim [0, p(u/u)] \)
\( \sim [0, p(j/j)] \)

\( \emptyset \rightarrow \cdot \rightarrow \) word \( j \emptyset \)

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

\( \sim [0, p(j/j)] \)

\( \emptyset/u \)
\( j/d \)
\( j/k \)
\( d \)

(van Gael+al ‘08), (Huggins+Wood ‘12)
Making this work in practice

**Different cutoffs**
- Separate cutoffs for letter and word transitions
- Letter cutoffs critical in discarding bad hypotheses
- Can’t be *too* different: introduces bias!

**The infinite prior**
- Prior over words is infinite: so is FST!
- Original paper uses sampling to deal with this: not efficient enough
- Treat prior as another FST...
  - But this introduces bias as well!
- Need to use Metropolis-Hastings rejection step (but usually accept)
Search strategies

Changing one utterance at a time does not collapse common variants:

\[ w\backslash t, \ w\backslash d \]

\[ ju, \ j\emptyset \]

Too many steps needed to convert all tokens...
Search strategies

Changing one utterance at a time does not collapse common variants:

\[ w\&t, \ w\&d \]
\[ ju, \ j\varnothing \]

Too many steps needed to convert all tokens...

**Phase of maximizing word sequence probabilities**

- Using two different annealing rates
- Rates \( >> 1 \) for word sequence maximize LM probs
- Overgeneralizes lexical items...
- Bad mergers usually unmerge when phase ends
Developmental speculation

System temporarily overgeneralizes words

- Group ðis, ðat, ðey
- Or hypothesize inserted/deleted segments: ɛn and ɛniʃ
- Short, vowel-heavy words particularly vulnerable

Evidence from development?

- Don’t know any proposals of this theory
- *(Merriman+Schuster ‘99)*: 2-4 year olds think “japple” might mean “apple” under some circumstances
- Tomasello and others: children learn multiword “chunks”
- Can these be reinterpreted as evidence for phonetic overgeneralization?

Perhaps can test via new experiments...
### Preliminary experiment

#### 1000 line dataset

<table>
<thead>
<tr>
<th>Tokens (boundaries only)</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>No channel</td>
<td>56</td>
<td>69</td>
<td>62</td>
</tr>
<tr>
<td>Joint</td>
<td>64</td>
<td>69</td>
<td>66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tokens (bounds and forms)</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>No channel</td>
<td>40</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>Joint</td>
<td>50</td>
<td>54</td>
<td>52</td>
</tr>
</tbody>
</table>

### Initial finding

Model with channel is better segmenter
- Better precision, fewer breaks overall
- Much better at predicting intended forms
  - Reassuring but not really surprising
Conclusions

- Data with variations is problematic for models of early lexical acquisition
- Possible to learn phonetics jointly with LM
- Learning synergy improves performance
- Seems possible to do everything jointly...
  - But requires some constraints in learning
Implications and future work

**Getting the rest of the way to acoustics will be tricky**

- Perhaps fully joint model like *(Feldman+al ‘09)*?
- Or pre-clustering like *(Varadarajan+al ‘08)*?

Probably some hidden surprises... results here show variation can be very problematic!
Implications and future work

Getting the rest of the way to acoustics will be tricky

- Perhaps fully joint model like (Feldman+al ‘09)?
- Or pre-clustering like (Varadarajan+al ‘08)?

Probably some hidden surprises... results here show variation can be very problematic!

Mechanisms for inference require some constraints

- The number of hypotheses our learner considers is vast...
- Keeping it manageable requires multiple interacting random filters

More study needed to find what infants are doing

Thanks

Mary Beckman, Laura Wagner and Lacqueys; Eric Fosler-Lussier, William Schuler and Clippers; funded by EPSRC; thanks for listening!