Bootstrapping a Unified Model of Lexical and Phonetic Acquisition

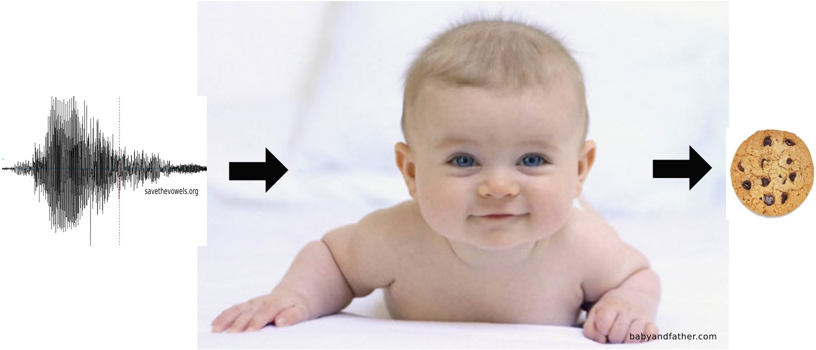
Micha Elsner Sharon Goldwater Jacob Eisenstein

School of Informatics
University of Edinburgh

School of Interactive Technology
Georgia Institute of Technology

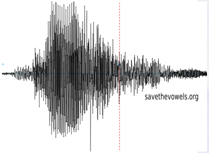
July 9, 2012

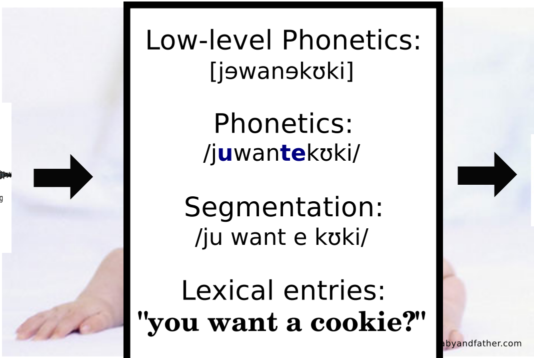
Early language learning



"you want a cookie?"

Early language learning





Low-level Phonetics:
[jəwanəkʊki]

Phonetics:
/j**u**wan**te**kʊki/

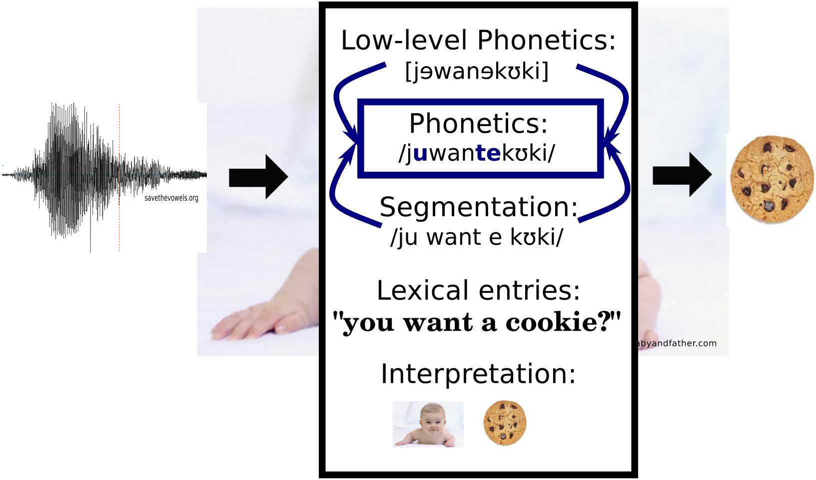
Segmentation:
/ju want e kʊki/

Lexical entries:
"you want a cookie?"

Interpretation:

Early language learning



Pronunciations vary

Variation

"Canonical" /want/ ends up as [wan] or [wãʔ]

Causes of variation

- Coarticulation (want ðə vs wãʔ wʌn)
- Prosody and stress (ði vs ðə)
- Speech rate
- Dialect

Learning sounds, learning words

How do infants learn that $[\text{jə}]$ is really $/\text{ju}/$?

Pipeline model

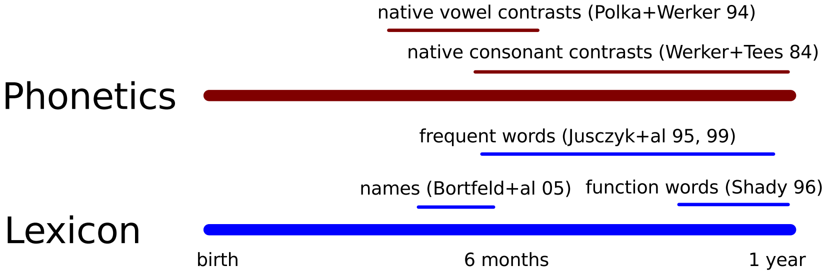
- Infant learns English phonetics/phonology first...
- "Unstressed vowels reduce to [ə]!"
- ...then learns the words

Joint model

(Feldman+al '09), (Martin+al forthcoming)

- Hypotheses about words support hypotheses about sounds...
- And vice versa
- "If [jə] is the same as [ju], perhaps vowels reduce!"

Developmental evidence supports joint model



following presentations by Feldman 09, Dupoux 09

Key developments at roughly the same time

This paper

Learn about phonetics and lexicon

Given low-level transcription with word boundaries:

$[\text{jə wã}ʔ\ \text{wʌn}]$

Infer an *intended* form for each surface form:

$/\text{ju want wʌn}/$

Inducing a language model over intended forms:

$p(/\text{want}/ \mid /\text{ju}/)$

And an explicit model of phonetic variation:

$p(/\text{u}/ \rightarrow [\text{ə}])$

Previous work

Learn about the lexicon

Segment words from intended forms (no phonetics):

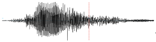
/juwantwʌn/ → /ju want wʌn/

(Brent '99, Venkataraman '01, Goldwater '09, many others)

Segment words from phones (no explicit phonetics or lexicon):

(Fleck '08, Rytting '07, Daland+al '10)

Word-like units from acoustics (no phonetic learning or LM):

 → want

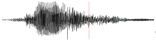
(Park+al '08, Aimetti '09, Jansen+al '10)

Previous work

Learn about the lexicon

Learn about phonetics

Discover phone-like units from acoustics (no lexicon):

 $\longrightarrow$ [u]

(Vallabha+al ‘07, Varadarajan+al ‘08, Dupoux+al ‘11, Lee+Glass here!)

Previous work

Learn about the lexicon

Learn about phonetics

Learn both

Supervised: (speech recognition)
Tiny datasets: (Driesen+al '09, Rasanen '11)
Only unigrams/vowels: (Feldman+al '09)

Previous work

Learn about the lexicon

Learn about phonetics

Learn both

Us

No acoustics, but...
Explicit phonetics and language model...
Large dataset

Overview

Motivation

Generative model

- Bayesian language model + noisy channel
- Channel model: transducer with articulatory features

Inference

- Bootstrapping
- Greedy scheme

Experiments

- Data with (semi)-realistic variations
- Performance with gold word boundaries
- Performance with induced word boundaries

Conclusion

Overview

- Motivation
- Generative model
  - Bayesian language model + noisy channel
  - Channel model: transducer with articulatory features
- Inference
  - Bootstrapping
  - Greedy scheme
- Experiments
  - Data with (semi)-realistic variations
  - Performance with gold word boundaries
  - Performance with induced word boundaries
- Conclusion

Noisy channel setup

loosely based on Goldwater+al 09

random lexicon

want, ju...

word-to-word transition probabilities

p(want|ju), p(to|want)

intended utterances

ju want wan
want e kʊki

our contribution

noisy channel

character sequence rewrite probabilities

p(u → ə : j_$)

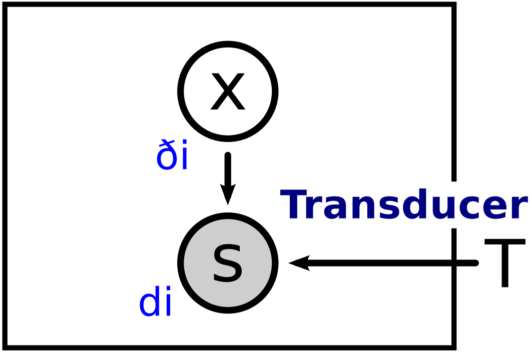
surface (observed)

jə wa? wʌn
wan ə kʊki

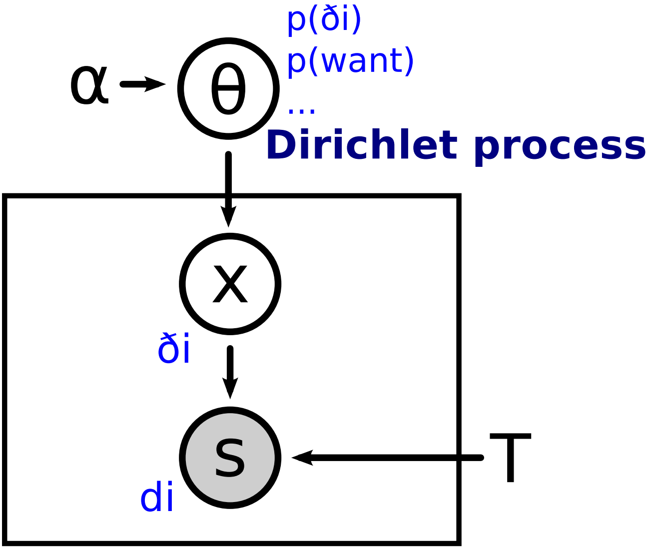
Graphical model

Presented as Bayesian model to emphasize similarities with (Goldwater+al '09)

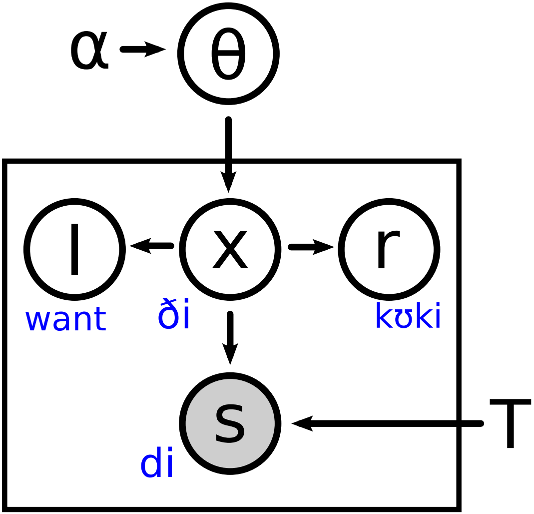
- Our inference method approximate



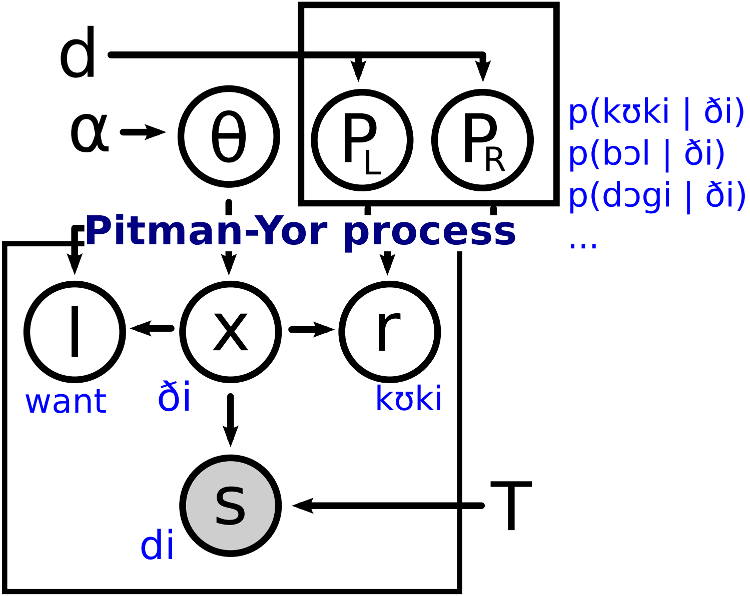
Graphical model



Graphical model



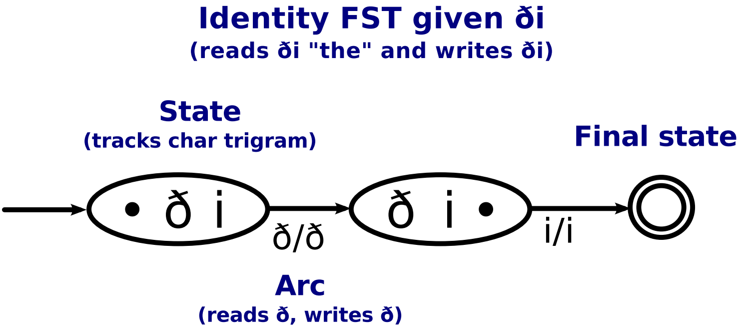
Graphical model



Transducers

Weighted Finite-State Transducer

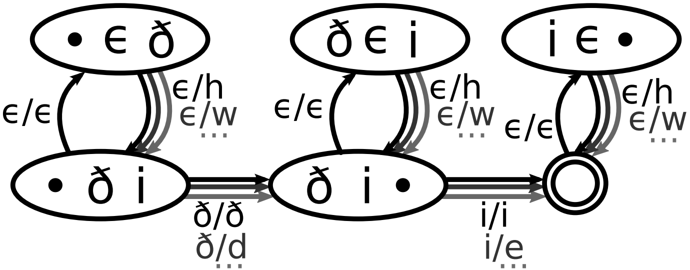
Reads an input string
Stochastically produces an output string
Distribution $p(out|in)$ is a hidden Markov model



Our transducer

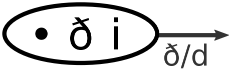
Produces any output given its input
Allows insertions/deletions

Reads ði, writes anything
(Likely outputs depend on parameters)



Probability of an arc

How probable is an arc?



Log-linear model

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

- Score of arc $\propto exp(w \cdot f)$

following (Dreyer+Eisner ‘08)

Articulatory features

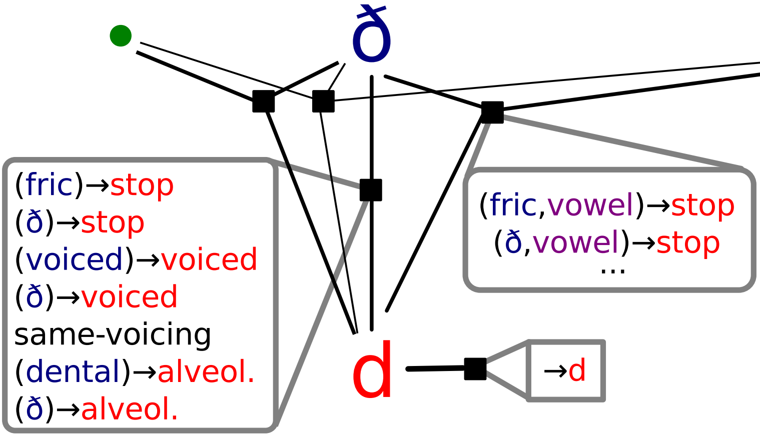
- Represent sounds by how produced
- Similar sounds, similar features
  - ð: voiced dental fricative
  - d: voiced alveolar stop

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:



Learned probabilities

- ð i $\longrightarrow$

ð	.7
n	.13
θ	.04
d	.02
z	.02
s	.01
ϵ	.01
$\ldots$	$\ldots$

Overview

- Motivation
- Generative model
  - Bayesian language model + noisy channel
  - Channel model: transducer with articulatory features
- Inference
  - Bootstrapping
  - Greedy scheme
- Experiments
  - Data with (semi)-realistic variations
  - Performance with gold word boundaries
  - Performance with induced word boundaries
- Conclusion

Inference

Bootstrapping

Initialize: surface type → itself ([di] → [di])

Alternate:

- Greedily merge pairs of word types
  - ex. intended form for all [di] → [ði]
- Reestimate transducer

Inference

Bootstrapping

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

Alternate:

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

Greedy merging step

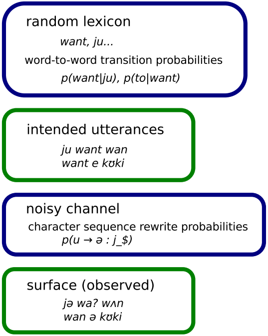
Relies on a **score** Δ for each pair:

- $\Delta(u, v)$: approximate change in model posterior probability from merging $u \rightarrow v$
- Merge pairs in approximate order of Δ

Computing Δ

$\Delta(u, v)$: 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)



Review

Bootstrapping

Alternate:

- Greedily merge pairs of word types
  - Based on Δ
- Reestimate transducer
  - Using Viterbi intended forms from merge phase
  - Standard max-ent model estimation

Overview

Motivation

Generative model
- Bayesian language model + noisy channel
- Channel model: transducer with articulatory features

Inference
- Bootstrapping
- Greedy scheme

Experiments
- Data with (semi)-realistic variations
- Performance with gold word boundaries
- Performance with induced word boundaries

Conclusion

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

- {ði, di, ðə} cluster can be identified by any of these

Score by tokens and types (lexicon).

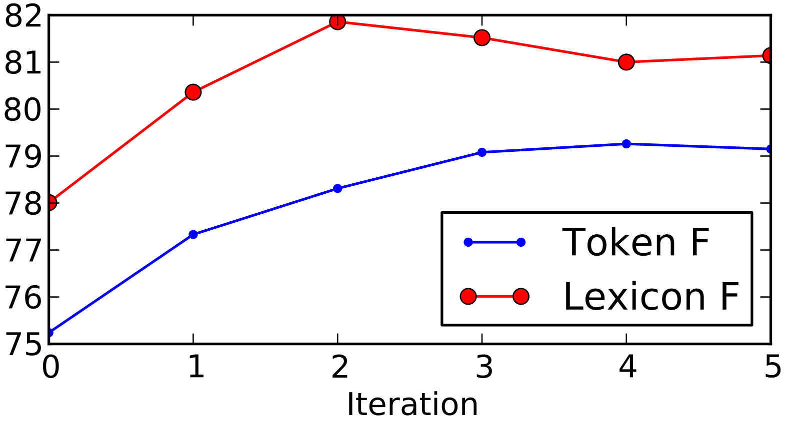
With gold segment boundaries

Scores (correct forms)

	Token F	**Lexicon (Type) F**
Baseline (init)	65	67
Unigrams only	75	76
Full system	**79**	**87**
Upper bound	91	97

Learning

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



Induced word boundaries

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

Scores (correct boundaries and forms)

	Token F	Lexicon (Type) F
Baseline (init)	44	43
Full system	**49**	**46**

After clustering, remove boundaries and resegment: sadly, no improvement

Conclusions

- Models of lexical acquisition must deal with phonetic variability
- First to learn phonetics and LM from naturalistic corpus
- Joint learning of lexicon and phonetics helps

Future Work

- Better inference
  - Token level MCMC/joint segmentation (in progress!)
- Real acoustics
  - Removes need for synthetic data