Taking the child's view:

Syllable-based Bayesian inference as a (more) plausible word segmentation strategy

Lawrence Phillips

Lisa Pearl

UC Irvine

Word Segmentation: Outline

- Infant representation: syllables vs. phonemes
- Incorporate cognitive constraints
- Discover evidence for ”Less is More” (Newport 1990)
  - Less-optimal learners perform better
- The unit of representation is extremely crucial to our interpretation of results

Word Segmentation

- Infants begin segmenting words out of fluent speech by 7.5 months (Jusczyk et al. 1999)
  - Stress Patterns: 9 months (Echols et al. 1997)
  - Phonotactics: 9 months (Jusczyk et al. 1993)
  - Phonemes: 10-12 months (Werker & Tees 1984)
- Word Segmentation is a foundation of later linguistic knowledge

Word Segmentation

- One popular explanation for how infants learn to segment words is from distributional information
- One basic form of distributional information which we know children have access to is Transitional Probabitities (TPs: Saffran et al. 1996; Pelucchi et al. 2009)

Modeling Word Segmentation

- Transitional Probabilities (TPs)

  ha → ppy → ki → tty

  H L H

  - Find word boundaries as TP-minima

  - But fails for monosyllabic sequences (Yang 2004; Gambell & Yang 2006)

    look → at → the → dog

    L L L

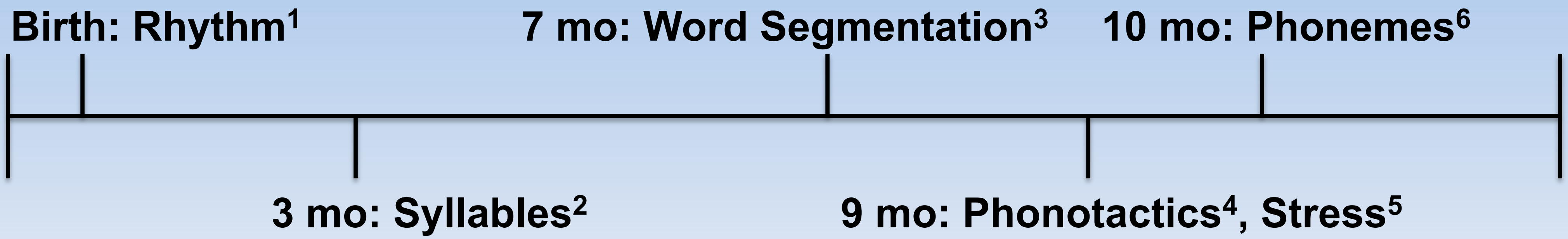
Assumptions in Word Segmentation

- Bayesian Modeling using TPs
  - Goldwater, Griffiths, Johnson (GGJ; 2009)
    - Builds a lexicon
    - Tracks TPs over phonemes
  - Pearl, Goldwater, Steyvers (PGS; 2010,2011)
    - Update GGJ to include cognitive constraints
    - Find a limited ”Less is More” effect

Bayesian Word Segmentation

- Bayesian models of Word Segmentation (GGJ succeed by tracking TPs while building a lexicon
  - Implicit bias for small lexicon (group together commonly occuring units)
  - Implicit bias for shorter words (don't group too much!)
- These models succeed, but...
  - Assume knowledge of phonemes

Speech Perception: 1st Year



- Begin with ***global*** perception
  - Rhythm, # of syllables
- Gain more ***specific*** representations
  - Syllables, phonemes, stress

[1]Nazzi et al. 1998
[2]Eimas 1999
[3]Jusczyk et al. 1999
[4]Echols et al. 1997
[5]Jusczyk et al. 1993
[6]Werker & Tees 1984

Phoneme Acquisition

- Phoneme Acquisition (~10 months) comes ***after*** Word Segmentation (~ 7 months)
- What other units do children use to represent language?
  - **Syllables** (~ 3 months (Eimas 1999))

    happykitty = ha / ppy / ki / tty
- How does word segmentation occur before phonemes are known?
- What role does this assumption play?

Syllabic Bayesian Modeling

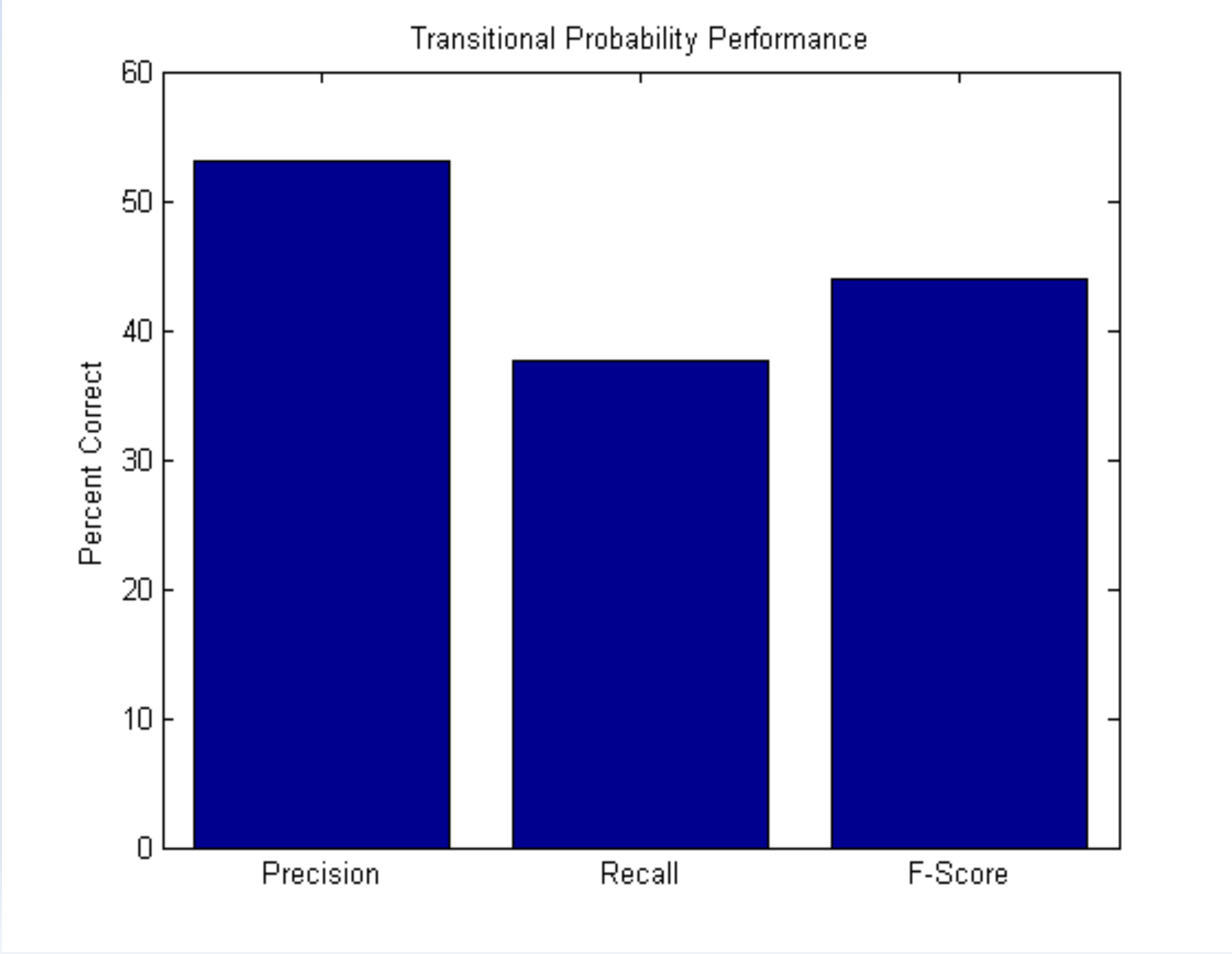
- Adapt previously successful Bayesian models (PGS, GGJ) to treat syllables as basic unit
  - Simplifies task: Fewer possible boundaries
  - But: ~40 phonemes, ~4000 syllables
- Syllabify Pearl-Brent corpus (MacWhinney 2000)
  - Child-directed speech (< 9 months)
  - 28,391 utterances, average 3.4 words/utterance
- Based on human judgments and Maximum-Onset Principle

Analysis

- We investigated both *Unigram* and ***Bigram*** models
  - Unigram: Words appear independently
  - ***Bigram***: Any word depends on the word before it
- We measure performance on ***Word Tokens*** as opposed to boundaries or lexical items
- We have 3 measures
  - Precision: # correct / # guessed
  - Recall: # correct / # true
  - ***F-Score***: Harmonic mean = (2 * P * R) / (P + R)

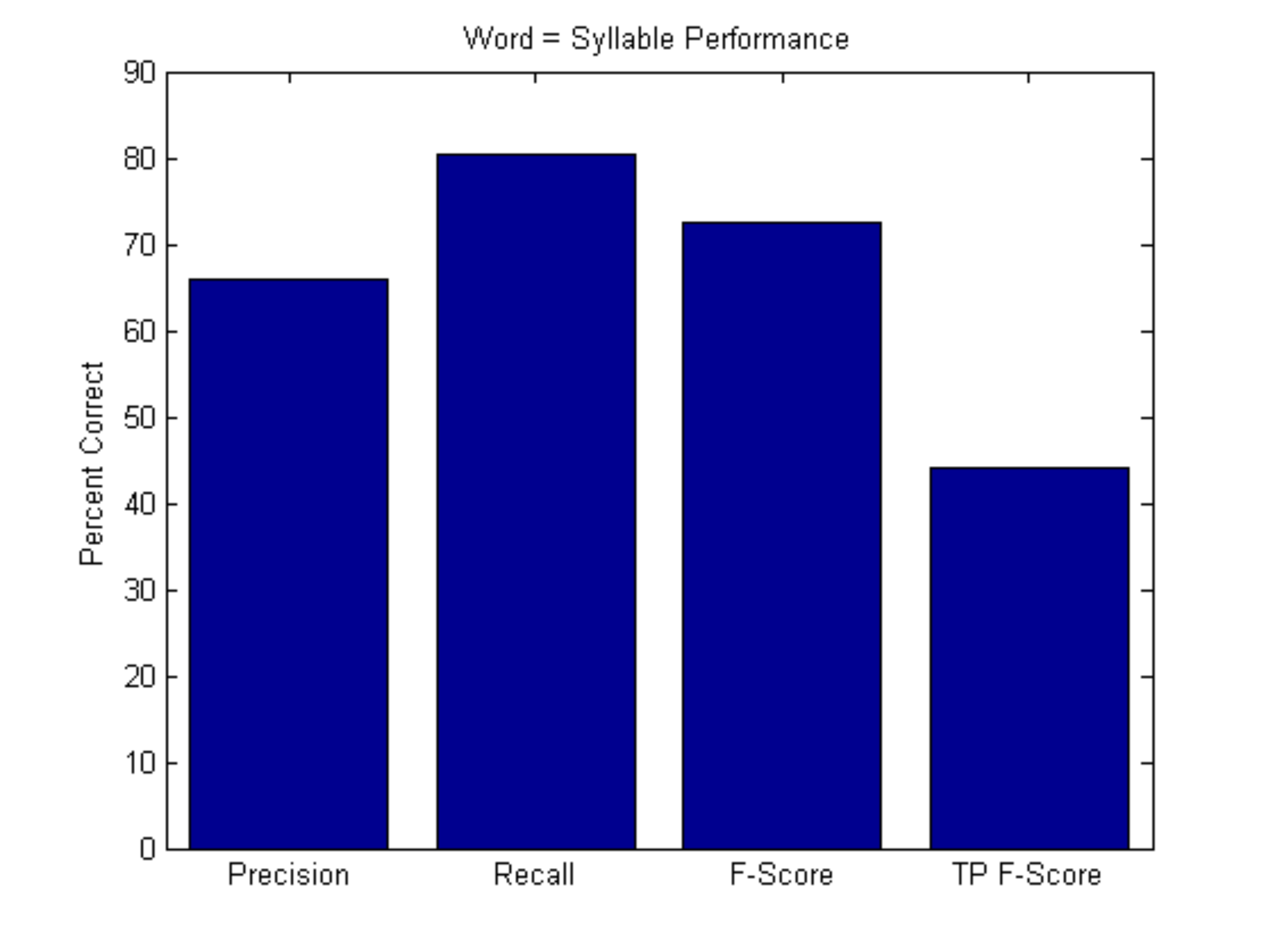
Other Syllable Models

- Transitional Probability model
  - Saffran et al. (1996) that children track TPs over syllables
  - undersegmentation



Other Syllable Models

- Syllable = Word
  - Doesn't match human performance (oversegmentation)



Other Syllable Models

- Gambell & Yang (2006), Yang & Lignos (2010)
  - Heuristic Models of Word Segmentation
  - Models require Unique Stress Constraint (USC)
    - 1 word = max. 1 primary stress
- Bayesian modeling
  - Doesn't require USC
  - More powerful than previously applied purely distributional models

PGS Models

	TP	Syl = Word	Batch Ideal
Token F-score	43.98	72.41	76.65

- Batch Ideal learner

  (GGJ 2009: Markov Chain Monte Carlo)

  - Sees all data at once
  - Remembers every decision, has unlimited computational resources
  - Uses Gibbs sampling, hierarchical Dirichlet Process

PGS Models

	TP	Syl = Word	Batch Ideal	DPM
Token F-score	43.98	72.41	76.65	74.46

- Online Ideal learner

  (**DPM**: Dynamic Programming with Maximization)

  - Processes each utterance in sequence
  - Chooses most optimal segmentation, remembers all decisions
  - Uses Viterbi algorithm to compute highest probability segmentation, given previous utterances

PGS Models

	TP	Syl = Word	Batch Ideal	DPM	DPS
Token F-score	43.98	72.41	76.65	74.46	76.70

- Online Sub-optimal learner

  (**DPS**: Dynamic Programming with Sampling)

  - Chooses segmentation probabilistically
  - Remembers all decisions
  - Uses Forward algorithm to compute probabilities and chooses based on each segmentation's likelihood

PGS Models

	TP	Syl = Word	Batch Ideal	DPM	DPS	DMCMC
Token F-score	43.98	72.41	76.65	74.46	76.70	86.19

- Online Memory-constrained learner

  (**DMCMC**: Decayed Markov Chain Monte Carlo)

  - Tends to ”remember” only recent decisions
  - Implemented with Decayed Markov Chain Monte Carlo (Marthi et al. 2002), choosing word boundaries to sample based on a decaying function

Results

- Memory-constrained learners outperform an ”optimal” Bayesian learner.
- Online algorithms have many benefits over batch processes (Liang & Klein 2009)
  - Avoid local minima, quick convergence
- …BUT we see ***decreased*** performance for our online optimal model!
- Sub-”optimal” segmentation, particularly memory constraints aid in learning to segment words

Less is More

- These findings support a view of language learning: The ”Less is More” hypothesis
  - Limited memory and cognitive resources help in learning language
  - ”Less is More” applies to adult language learners (Chin & Kersten 2010; Kersten & Earles 2001; Cochran et al. 1999)
  - Here: computational support for this phenomenon in word segmentation

Less is More

- PGS also found results for ”Less is More” but mostly for Unigram DPM & DMCMC models
  - Potentially based on ”online” advantage (Liang & Klein 2009)
- By changing the underlying unit of representation we can see this pattern of results much more clearly
- Unit of representation clearly matters for how we interpret our results

Open Questions

- What role does syllable type or syllabification method play in our results?
  - Run model over infant-directed speech in German (many syllable types) and Spanish (fewer syllable types)
- Incorporate knowledge of predominant stress patterns
  - Infants segment words at 7.5 months easier if they follow the predominant stress pattern of the language (Jusczyk et al. 1999)

Thanks

- Galia Barsever, Caroline Wagenaar, Jim White
- Mark Johnson and Sharon Goldwater
- Iain Murray, Alex Ihler
- Everyone at IPAM Summer Institute 2011
- Our Reviewers

Results

Unigram Models	TP	TR	**TF**	BP	BR	**BF**	LP	LR	**LF**
Ideal	65.34	45.85	**53.89**	92.20	56.38	**71.63**	45.59	71.78	**55.75**
DPM	71.97	48.58	**57.96**	98.07	52.50	**68.32**	37.35	53.14	**43.86**
DPS	74.33	53.27	**62.03**	97.20	57.90	**72.51**	41.17	57.21	**47.87**
DMCMC	67.31	49.67	**57.16**	96.82	60.55	**74.48**	48.74	72.79	**58.38**
Bigram Models									
Ideal	81.84	72.08	**76.65**	96.05	79.67	**87.09**	65.27	79.06	**71.50**
DPM	81.49	68.57	**74.46**	96.67	74.84	**84.35**	56.96	70.46	**62.99**
DPS	82.96	71.34	**76.70**	96.48	77.20	**85.75**	57.83	71.23	**63.83**
DMCMC	86.19	85.23	**86.19**	94.01	91.05	**92.49**	74.18	77.28	**75.70**
Comparison Models									
TransProb	53.03	37.57	**43.98**	90.00	53.14	**66.82**	11.72	63.08	**19.77**
Syl = Word	65.89	80.37	**72.41**	76.26	100	**86.53**	59.79	43.25	**50.19**

Eimas 1999

- Investigated 3- to 4-month olds ability to form categorical representations of consonants and syllables
- Tested infants on CV and CVC utterances
  - No categorical representation of initial consonant
- Tested infants on bisyllabic utterances
  - Strong categorical representation of initial syllables
  - Weak representation of final syllables

GGJ 2009

$P(\text{utterance}) = \prod(P(\text{word}_i)[1\text{-}P(\text{end of utt})]) * P(\text{final word})P(\text{end of utt})$

1) Decide if w_i is a novel lexical item

2) a. If so, generate a phonemic form

b. If not, choose an existing lexical item

$P(w_i \text{ is novel}) = \alpha / (n + \alpha)$

$P(w_i = x_i \ldots x_M \mid \text{novel}) = P_{\#} (1\text{-}P_{\#})^{M-1} \prod P(x_j)$

$P(w_i = l \mid \text{not novel}) = \#\text{ of } l\text{'s} / \#\text{ of words}$