Using Prosody to Bootstrap Word Segmentation in a More Realistic Learning Environment

Constantine Lignos and Charles Yang

University of Pennsylvania

BCCCD 11 Symposium: The Role Of Prosody In Guiding Language Learning In Pre-Lexical Infants

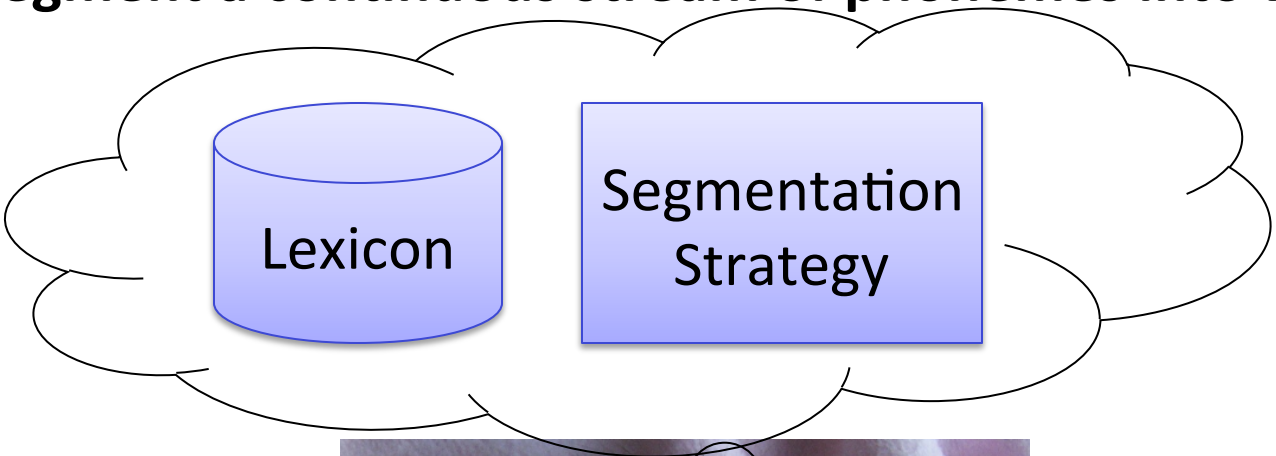
1/15/2011

Penn
University of Pennsylvania

I. Overview

The task: word segmentation

- **Segment a continuous stream of phonemes into words**



bigdrum
horse
whoisthat
playcheckers

big drum
horse
who is that
play checkers

A summary of our approach

- **The learner uses:**
  - A linguistic constraint (*Unique Stress Constraint*) to define the structural description of a word
  - A simple algebraic approach that uses an existing lexicon to identify more words
- **It operates on *syllables,* unlike most computational models which operate on phonemes**
- **It is a *bootstrapping learner*—it uses simple heuristics to get the first words and uses a segmentation strategy to learn new words**

What makes our approach different

- **More cognitively oriented than most models**
  - Aligned with the computations children can perform, as shown by experiments
  - Does not require intensive computational optimization
- **Assumes more about the learner's capabilities beyond identifying phonemes**
  - Able to learn syllabification through phonotactics (Onishi et al., 2002)
  - Able to map acoustic signal to strong/weak stress on syllables (Johnson & Jusczyk, 2001)
  - These assumptions are inline with developmental evidence, unlike many other computational models

Some relevant developmental studies

- **In brief, child learners appear to be able to:**
  - Operate on syllables from birth (Bijeljac-Babic et al., 1993)
  - More easily identify novel words at the beginning or ends of utterances at 8 months (Seidl & Johnson, 2006)
  - Identify syllabic stress, learn a stress pattern for their language and prefer it over transitional probability cues (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003)
  - Initially (7.5 months) rely heavily on a dominant stress pattern as a segmentation cue, later adapting to use multiple cues (10 months) and achieving adult-like competence at 24 months (Jusczyk et al., 1999)

Recent computational models

- **Other models have used Bayesian methods using transitional probabilities and lexicons (Brent, 1999; Goldwater et al., 2009; Johnson and Goldwater, 2009)**
  - While these models demonstrate interesting statistical techniques, their connection to cognition and infant learning is not clear
  - The techniques used show that when combined with sophisticated machine learning techniques, transitional probabilities can be used for segmentation in English
- **Our focus here is different, we show that with more informative cues and the right learning model a learner can succeed with an extremely simple approach**

II. Our algorithm

Overview

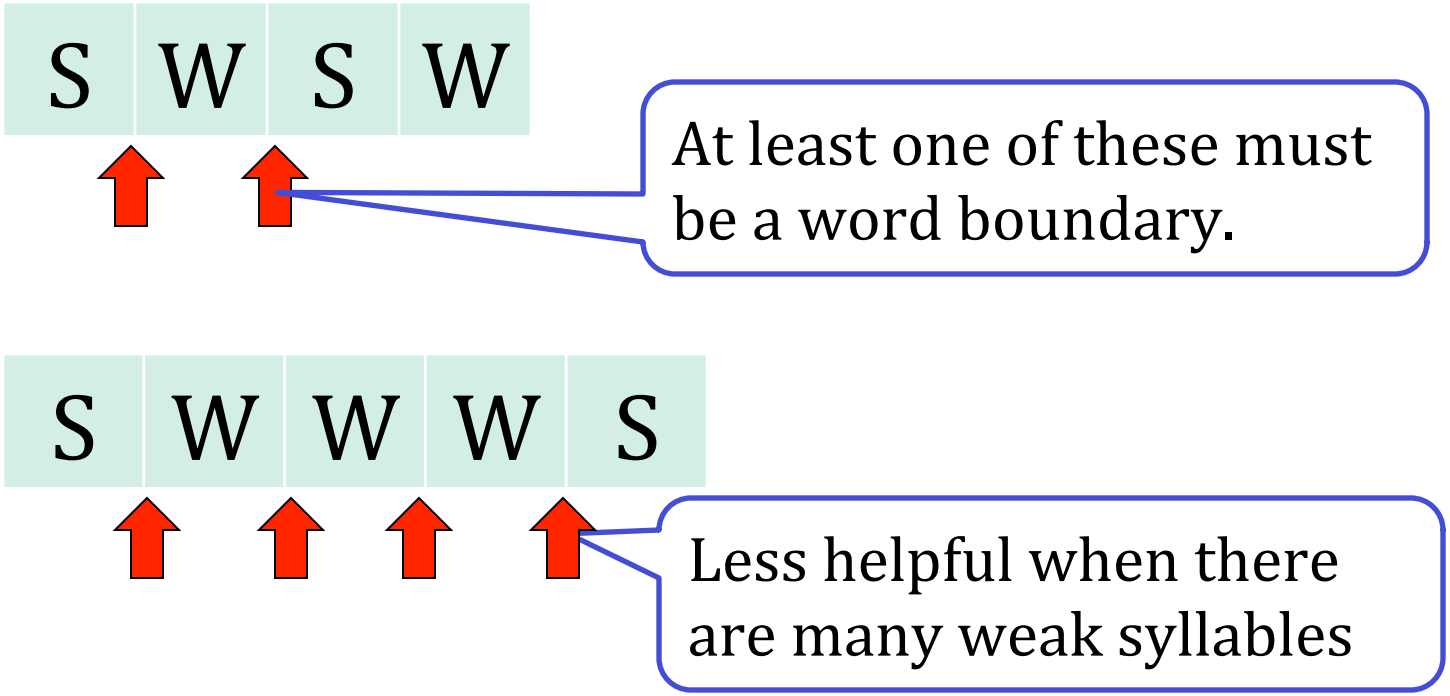
- **The segmenter has a *lexicon* of words it believes are in the language that it builds over time**
  - It starts empty, and words are added based on the words hypothesized in the segmentation of each utterance
- **The segmenter operates *online***
  - It segments one utterance at a time, and cannot remember previous utterances or how it segmented them
- **The segmenter works *left-to-right* on each utterance and inserts word boundaries**

Unique Stress Constraint (USC)

A word can bear at most one primary (strong) stress.

—(Halle & Vergnaud, 1987; Yang, 2004; Gambell & Yang, 2006)

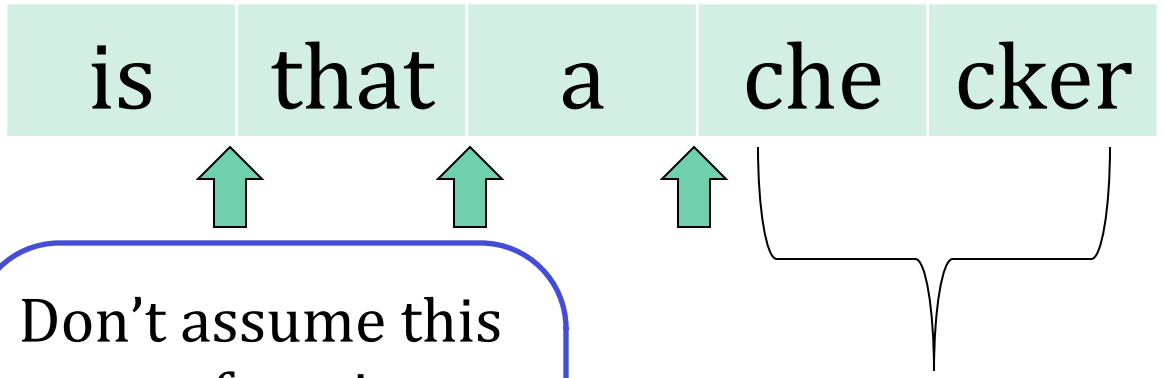
Assume we have strong (S) and weak (W) syllables, how can we use these to segment?



Subtractive Segmentation

- **We can use words we already know to break up the utterance:**

Lexicon:
a
is
that
checker

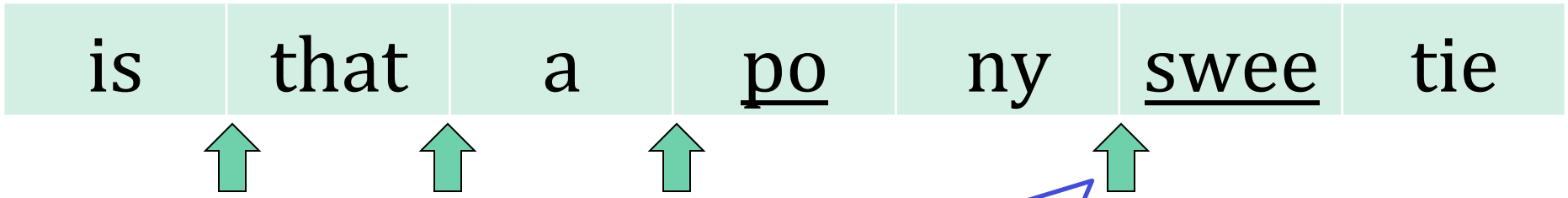


Don't assume this means function words are the first ones learned; we'll return to that shortly.

Treat the remainder as a word, add it to the lexicon

Subtractive Segmentation and USC

- **Work from left to right through the utterance**
  - Insert a word boundary where needed to prevent a word from having two strong stresses
  - If the current position starts with a word in the lexicon, segment it off



Lexicon:
a
is
that

Without metrical information, the segmenter would not place a boundary here.

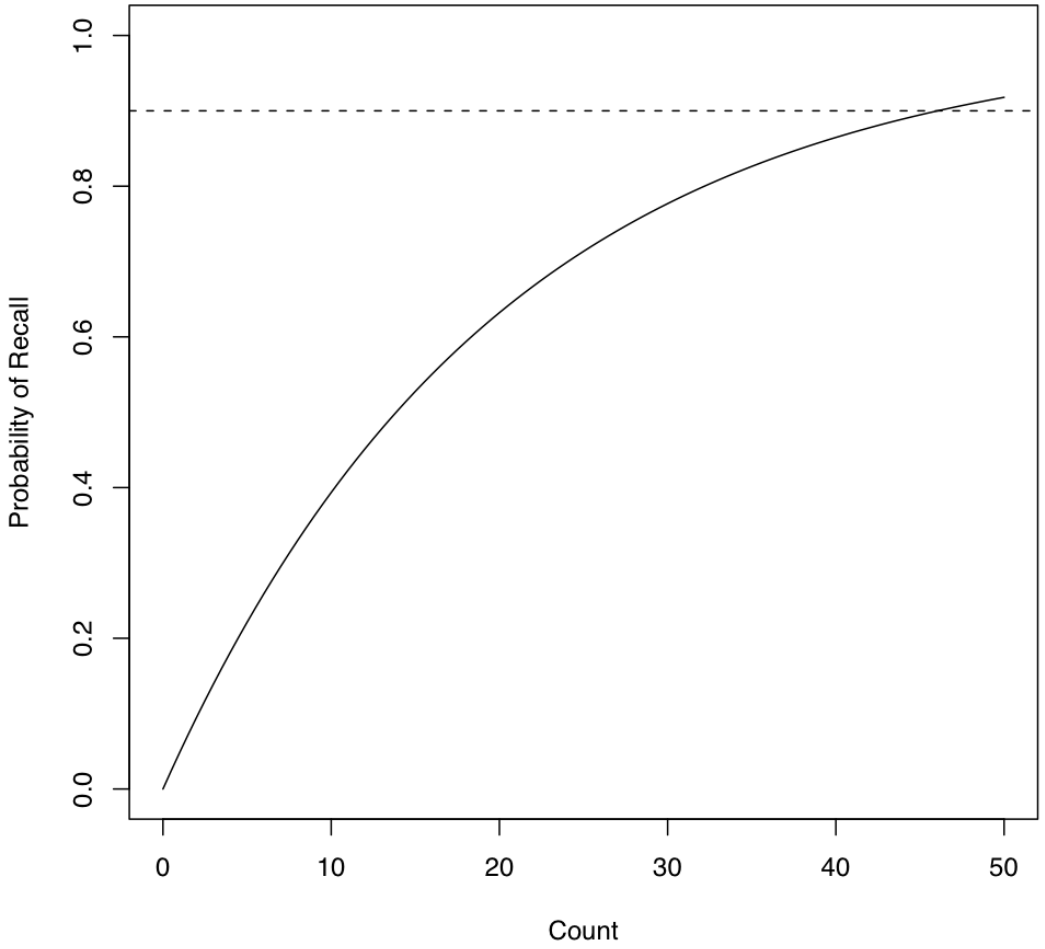
Probabilistic memory

- **It's unlikely that infants can always remember every word after just one first hearing or correctly identify the syllables every time**
  - Since subtraction relies on recalling words, a simple set model of what's "in" the lexicon may be unfair
- **Let the probability of recalling a word grow with the number of times it is hypothesized**
- **Our probability function:**

$$p_r(word) = 1.0 - e^{-\alpha c(word)}$$

  - —$p_r(word)$- probability of recalling a word
  - —$c(word)$- number of times word has been hypothesized before
  - —α- constant, fixed at 0.05 in our experiments

Probabilistic memory function



III. Analysis

Our evaluation corpus

- **Constructed from the Brown (1973) subset of CHILDES English (Adam, Eve, Sarah)**
- **Pronunciations and stress for each word come from CMUDICT**
- **Syllabified using *Maximize Onset* with a list of valid consonant clusters of English**

Einstein	*Input*
AY.N.S.T.AY.N	*Pron. Lookup*
AY.N|S.T.AY.N	*Syllabification*

- **Stress modified to better reflect natural speech**
  - No adjacent primary stresses (Liberman and Prince, 1977; Selkirk 1984)

Evaluation

- **Precision and recall calculated over the input corpus**
  - Precision: Percent of word boundaries the learner predicts that are correct
  - Recall: Percent of word boundaries in the gold segmentation that the learner predicts
  - F-score (F1): Harmonic mean of precision and recall
  - Undersegmentation results in high precision and low recall, and oversegmentation the opposite
- **Two stress conditions:**
  - No stress- No stress information
  - Lexical stress- Lexical stress information

Performance

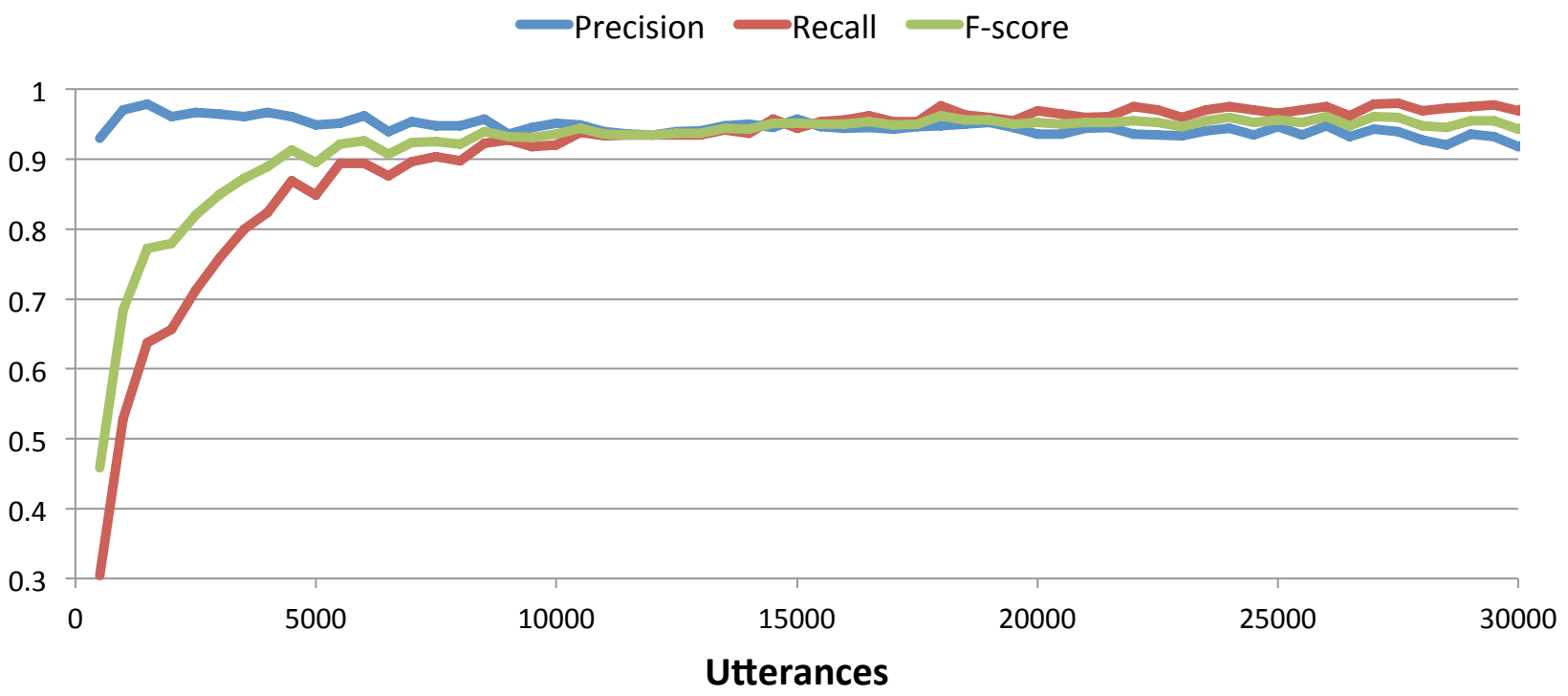
- **With or without stress information, our simple bootstrapping approach performs well**
- **Adding lexical stress information reduces the F1 error significantly**
  - Even with the reduced stress information available in our simulation, stress information gives the learner a significant advantage

Probabilistic Memory	Precision	Recall	F-score	Error Reduction
No Stress	92.18%	91.93%	92.06%	20.35%***
Lexical Stress	93.67%	93.57%	93.62%	

***p < 2.2e-16

Learning curve

- **Learner starts undersegmenting, as it learns achieves balance with slight oversegmentation**



Errors over time

- **Early:**
  - "Big drum" as "Bigdrum" [First utterance in corpus]
    - —Because the learner's lexicon is empty and there's only one primary stress in the utterance, no segmentation occurs
  - "How many trucks?" as "Howmany trucks?"
    - —Frequent function word collocations (is that, you are, what are, etc.) are often treated as one word
- **Late:**
  - "Want me to take it away from you" as "Want me to take it a way from you"
    - —Function word *a* mistakenly segmented off *away*, similar to *behave/be have* and *tulips/two lips* errors

How can the learner use its lexicon?

- **Our evaluation focuses on how well the learner segments the input, but the learner also builds a lexicon**
- **With a lexicon of reasonable quality, the learner can start to:**
  - Learn lexical stress pattern for the language
  - Learn distributions for in-word and between-word transitional probabilities
  - Learn the morphology of the language
- **Learning all of these things can aid word segmentation**

Conclusions and future work

- **Using lexical stress information can significantly aid the learner**
- **But the effectiveness of this approach needs to be demonstrated in other languages**
  - In languages without word-level stress, other cues will need to be used in tandem
- **We've focused on a single cue and subtractive segmentation, but adding other cues is likely to lead to better performance**
- **It would more interesting to get the stress from acoustic information**
  - Let me know if you're interested in doing this!

Thanks!

For more info, contact me at lignos@cis.upenn.edu