An explicit statistical model of learning lexical segmentation using multiple cues

Çağrı Çöltekin
University of Groningen
c.coltekin@rug.nl

John Nerbonne
University of Groningen
j.nerbonne@rug.nl

Abstract

This paper presents an unsupervised and incremental model of learning segmentation that combines multiple cues whose use by children and adults were attested by experimental studies. The cues we exploit in this study are *predictability statistics*, *phonotactics*, *lexical stress* and partial *lexical information*. The performance of the model presented in this paper is competitive with the state-of-the-art segmentation models in the literature, while following the child language acquisition more faithfully. Besides the performance improvements over the similar models in the literature, the cues are combined in an explicit manner, allowing easier interpretation of what the model learns.

1 Introduction

Segmenting the continuous speech stream into lexical units is one of the challenges we face while listening to other speakers. For competent language users, probably the biggest aid in identifying the word boundaries is the knowledge of the words. Not surprisingly, the models of adult word recognition depend heavily on a lexicon (see Dahan and Magnuson, 2006, for a recent review). The same can be observed in speech and language technology where all automatic speech recognition systems make use of a comprehensive lexicon.

Even with a comprehensive lexicon and an error-free representation of the acoustic input, the problem is not trivial, since the input is often compatible with multiple segmentations spanning the complete utterance. The problem, however, is even more difficult for a learner who starts with no lexicon. Fortunately, the lexicon is not the only aid for segmentation. Experimental research within last two decades has revealed an array of cues that are used by adults and children for lexical segmentation. These cues include, but are not limited to, *lexical stress* (Cutler and Butterfield, 1992; Jusczyk, Houston, et al., 1999), *phonotactics* (Jusczyk, Cutler, et al., 1993), *predictability statistics* (Saffran et al., 1996), *allophonic differences* (Jusczyk, Hohne, et al., 1999), *coarticulation* (E. K. Johnson and Jusczyk, 2001), and *vowel harmony* (Suomi et al., 1997). The relative utility or dominance of these cues is a matter of current debate. However, it seems uncontroversial that none of these cues solves the segmentation problem alone and, when available, they are used in conjunction.

Along with experimental research on segmentation, a large number of computational models have been proposed in the literature. The early studies typically made use of connectionist models (e.g., Elman, 1990; Christiansen et al., 1998). Of these studies, Christiansen et al. (1998) is particularly interesting for the present study since it incorporates most of the cues used in this study. Using a simple recurrent network (SRN, Elman, 1990), Christiansen et al. (1998) demonstrated the usefulness of lexical stress, predictability statistics (included implicitly in any SRN model), and utterance boundaries, and showed that combining the cues improves the performance. The connectionist models have been instrumental in investigating a large number of cognitive phenomena. However, they have also been subject to the criticism that what a connectionist model learns is rather difficult to interpret. Furthermore, the performance achieved using connectionist models is far lower than that is expected from humans.

Models that use explicit representations in combination with statistical procedures (e.g., Brent and Cartwright, 1996; Brent, 1999; Venkataraman, 2001; Goldwater et al., 2009; M. Johnson and Goldwater, 2009) avoid both problems: these models perform better, and it is easier to reason about what they learn. Although these models were also instrumental in our understanding of the problem, they lack at least two aspects of connectionist models that fit human processing better. First, even though we know that human segmentation is incremental and predictive, most of these models process their input either in a batch fashion, or they require the complete utterance to be presented before attempting to segment the input. Second, it is generally difficult to incorporate arbitrary cues into most of these models.

Models that use explicit representations with incremental models exist (e.g., Monaghan and Christiansen, 2010; Lignos, 2011), but are rather rare. Furthermore, the investigation of cues and cue combination in segmentation is also relatively scarce within the recent studies (exceptions include the investigation of various suprevised models by Jarosz and J. A. Johnson, 2013).

The present paper introduces a strictly incremental, unsupervised method for learning segmentation where the learning method and internal representations are explicitly defined. Crucially, we use a set of cues demonstrated to be used by humans in solving the segmentation problem. The simulations results that we present are based on the same child-directed speech input used by many other studies in the literature.

The rest of this article is organized as follows: in the next section, we present a method for combining cues. Section 3 describes the cues used in this study. The simulations are described and results are presented in Section 4. A general discussion of the modeling framework and the simulation results are given in Section 5.

2 A cue combination method

We know that there is no single cue that always gives the correct answer in the lexical segmentation task. We also know that humans combine multiple cues when available. In this section we define a method to segment a given utterance using multiple boundary indicators, or cues, and learn to segment better by estimating usefulness of each indicator. In essence, each indicator makes a decision on each potential boundary location. The method combines these indicators' decisions to arrive at a hopefully more accurate decision. In machine learning terms, we formulate a number of binary classifiers, and aim to get a better classifier using a combination of them. This problems is a relatively well-studied subject in the machine learning literature (e.g., Bishop, 2006, chapter 14). Here a simple and well-known method, *majority voting*, will be used for combining multiple boundary indicators.

Majority voting is a common (and arguably effective) method in everyday social and political life. As a result, it has been well studied, and known to work well especially if each voter's decision is better than random on average, and votes are cast independently. In practice, even though the votes are almost never independent, majority voting is still an effective way of combining multiple classifiers (see Narasimhamurthy, 2005, for a discussion of the effectiveness of the method).

The majority voting combines each vote equally. Even though this may be a virtue in the social and political context, it is a shortcoming for a computational procedure that incorporates information from multiple sources with varying usefulness. We will use a simple augmentation of majority voting to model, *weighted majority voting* (Littlestone and Warmuth, 1994), that weighs the utility of the information provided by each source.

In weighted majority voting, the voters that make fewer errors get higher weights. In an unsupervised setting as ours, we do not know for certain when a voter makes an mistake. Instead, we take a voter's decision to be correct if it agrees with the majority. Initially we set all the weights to 1, trusting all the voters equally. We adopt an incremental version of the algorithm, where we keep the count of 'errors' made by each voter i, e_i, which is incremented every time the voter disagrees with the majority. After every boundary decision, first, the error counts are updated for each voter. Then, the weight, w_i, of each voter is updated using,

$$w_i \leftarrow 2\left(0.5 - \frac{e_i}{N}\right)$$

where N is the number of boundary decisions made so far, including the current one.

This update rule sets the weight of a voter that is half the time wrong (a voter that votes at random) to zero, eliminating the incompetent voters. If the votes of a voter are in accordance with the weighted majority decision almost all the time, the weight stays close to one.

3 Cues and boundary indicators

The combination method above allows us to combine an arbitrary number of boundary indicators. In our setting, each psychologically motivated cue is represented by multiple boundary indicators that differ based on the source of information used and the way this information is turned into a quantitative measure. This section introduces all of these cues, and the boundary indicators that stem from quantification of these cues in different ways.

3.1 Predictability statistics

At least as early as Harris (1955), it was known that a simple property of natural language utterances can aid identifying the lexical units that form an utterance: *predictability within the units is high, predictability between the units is low*. However, until the influential study by Saffran, Aslin, and Newport (1996), the idea was not investigated in developmental psycholinguistics as a possible source of information that children may use for segmentation. After Saffran et al. (1996) showed that 8-month-old infants make use of predictability statistics to extract word-like units from an artificial language stream, a large number of studies confirmed that predictability based strategies are used by adults and children for learning different aspects of language (e.g., Thiessen and Saffran, 2003; Newport and Aslin, 2004; Graf Estes et al., 2007; Thompson and Newport, 2007; Perruchet and Desaulty, 2008).

To use in our cue combination system, we need to quantify the notion of predictability. In this study, we use two information theoretic measures of predictability (or surprise), to define a set of boundary indicators. The first one, *pointwise mutual information* (MI) is defined as

$$\mathsf{MI}(l, r) = \log_2 \frac{P(l, r)}{P(l)P(r)}$$

where l and r are strings of phonemes to the left and right of the possible boundary location. We define our second measure, *boundary entropy* (H) of a potential boundary after string l as

$$\mathsf{H}(l) = -\sum_{r \in A} P(r|l) \log_2 (P(r|l))$$

where the sum ranges over all phonemes in the alphabet, A.[1]

The use of both the MI and the H is motivated by the finding that combination multiple predictability measures result in better segmentation (see Çöltekin, 2011, p.101, for an analysis). Furthermore, for asymmetric measures, like entropy, $\mathsf{H}(l)$ is clearly not the same as $\mathsf{H}(r)$. Motivated by the finding that children use 'reverse predictability' (Pelucchi et al., 2009), we also incorporate a reverse entropy measure in the present study.

In most studies in the literature, the context l and r are single basic units (phonemes in our case). The different phoneme context sizes may capture regularities that exist because of different linguistic units. The relation between the phoneme context size and the linguistic units, of course, is not clear-cut. However, for example, we expect context size of one to capture the regularities between the phonemes, while context size of two or three to capture regularities between larger units, such as syllables.

The above parameters result in an array of indicators. However, none of the indicators we use have a natural threshold to decide whether a given position is a boundary or not. To get a boundary decision out of a single measure (MI or H), we adopt a method similar to a commonly used unsupervised method that decides for a boundary at the 'peaks' of unpredictability. A particular shortcoming of this strategy, however, is that it can never find both boundaries of a single-phoneme word, as there cannot be two peaks one after another. To remedy this, the *partial-peak* strategy we employ here makes use of two sets of boundary indicators for each potential boundary: one posits a boundary *after* an increase in H (or a decrease in MI) and the other posits a boundary *before* a decrease in H.

3.2 Utterance boundaries

An attractive aspect of the predictability-based segmentation is that it does not require any lexical knowledge in advance—unlike other cues noted in Section 1. However, certain aspects of phonotactics, such as the regularities found at the beginning and end of words, can be induced from the boundaries already marked in the input without the need for a lexicon. As a result, clearly marked lexical unit boundaries may serve as another source of information that can bootstrap the acquisition of lexical units.[2]

[1]The input to children is better represented by 'segments' or 'phones'. However, since the data used in our simulations does not contain any phonetic variation, in this paper, we use the term phoneme when referring to the basic input unit.

[2]There are a number of acoustic cues (e.g., pauses) that are highly correlated with lexical unit boundaries. However, we do not make use of them in this study since they are considered to be unreliable, and they are not marked in the corpora at hand.

All models of segmentation in the literature use utterance boundaries implicitly by assuming that the words cannot straddle utterance boundaries. The explicit use of utterance boundaries to discover regularities about words is common in connectionist models (e.g., Aslin et al., 1996; Christiansen et al., 1998; Stoianov and Nerbonne, 2000). Similar use of utterance boundaries in non-connectionist models is rather rare. Three exceptions to this are the models described by Brent (1996), Fleck (2008) and Monaghan and Christiansen (2010). The method described in this section is similar to Fleck's method, where the model estimates the probability of observing a boundary given its left and right context, $P(b|l,r)$, where b represents boundary, and as before, l and r represent left and right contexts, respectively. If this probability is greater than 0.5, the model inserts a boundary. Using utterance boundaries and the pauses, Fleck (2008) presents a batch algorithm with a few ad hoc corrections that estimates the probabilities $P(b), P(l|b), P(r|b), P(l), P(r)$, and uses Bayesian inversion to estimate $P(b|l,r)$.

In this work, instead of $P(b|l,r)$, we estimate probabilities of utterance beginnings, $P(ub|r)$, and probabilities of utterance ends, $P(ub|l)$, where ub stands for utterance boundary. These probabilities can directly be estimated from the utterance edges in the input corpus, and can be used as cues for discovering non-initial or non-final boundaries. Similar to the predictability, using different length l and r we obtain a set of indicators for $P(ub|r)$ and $P(ub|l)$.

Unlike $P(b|l,r)$, for $P(ub|r)$ and $P(ub|l)$ we do not have a straightforward threshold to make a boundary decision. Instead, we appeal to the familiar solution, and use 'partial peaks' in these values as boundary indications.

3.3 Lexical stress

Lexical stress is one of the cues for segmentation that is well supported by psycholinguistic research (e.g., Cutler and Butterfield, 1992; Jusczyk, Houston, et al., 1999; Jusczyk, 1999). Lexical stress is used in many languages for marking the prominent syllable in a word. For languages that exhibit lexical stress, the prominent syllable will typically be in a particular position in the word, allowing discovery of the boundaries based on the position of stressed syllable.

Despite the prominence of stress as a cue for segmentation, there are relatively few computational studies that investigate use of stress. Christiansen et al. (1998) incorporates stress as a cue in their connectionist cue combination system. Swingley (2005) provides a careful analysis of stress patterns of the bisyllabic words found by a discovery procedure on mutual information and frequency. Gambell and Yang (2006) present surprisingly good segmentation results with a rule-based learner whose main source of information is lexical stress. One of the major problems with the these studies, which has also been carried over to the present study, is the lack of corpora with realistic stress assignment (see Section 4.1).

Our stress-based strategy is similar to the strategy used for learning phonotactics described in Section 3.2. Instead of collecting statistics about phoneme n-grams, we collect statistics over stress assignments on phoneme n-grams. However, the probabilities are estimated over already known lexical units. Given stress patterns l and r, we estimate $P(b|l)$ from endings of the known lexical units, and $P(b|r)$ from the beginnings of the lexical units. Again we use these quantities as indicators for variable length l and r. Using the partial-peak boundary decision strategy in combination with the weighted majority voting algorithm, as before, we define a set of boundary indicators and operationalize lexical stress as another cue for segmentation.

3.4 Lexicon

For adults, a comprehensive lexicon is probably the most useful cue for segmentation. We do not expect infants to have a lexicon at the beginning. However, as they build their lexicon, or 'proto-lexicon', they may put it in use for discovering novel lexical units. This is the main strategy behind the majority of state-of-the-art computational models of segmentation (e.g., Brent, 1999; Venkataraman, 2001; Goldwater et al., 2009). The models that guess boundaries rarely build and use an explicit lexicon (exceptions include Monaghan and Christiansen, 2010).

In this study we also experiment with an (admittedly naive) set of lexical cues to word boundaries. The idea is to indicate a boundary when there are word-like strings on both sides of the boundary candidate. In our usual majority voting framework, these form two additional sets of boundary indicators. First, given a possible boundary location, we simply count the frequencies of already known words beginning or ending at the position in question. The second indicator is based on the number of times the phoneme sequences surrounding the boundary found at the beginnings or ends of the previously discovered words. The second indicator is essentially the same as the phonotactics component discussed in Section 3.2, except that it is calculated using already known word types instead of utterance boundaries.

Similar to the other asymmetric indicators discussed previously, we have two flavors for each indicator. One indicating the existence of words to the right of the boundary candidate (words beginning at the boundary), and the other indicating the existence of words the left of the boundary candidate (word ending at the boundary). As with the other cues, these result in a set of indicators whose primary source of information is the potential lexical units in the learner's incomplete and noisy lexicon.

4 Experiments

4.1 Data

We use a child-directed speech corpus from the CHILDES database (MacWhinney and Snow, 1985). It was collected by Bernstein Ratner (1987) and the original orthographic transcription of the corpus was converted to a phonemic transcription by Brent and Cartwright (1996). The same corpus has been used by many recent studies. Following the convention in the literature the corpus will be called the *BR corpus*.

For the results reported for segmentation strategies that make use of lexical stress, the BR corpus was marked for lexical stress semi-automatically following the procedure described by Christiansen et al. (1998) for annotating the Korman corpus (Korman, 1984). The stress assignment is done according to stress patterns in the MRC psycholinguistic database. All single-syllable words are coded as having primary stress, and the words that were not found or did not have stress assignment in the MRC database were annotated manually.

4.2 Evaluation metrics

Two quantitative measures, *precision* (P), *recall* (R) and their harmonic mean *F_1-score* (F-score, or F, for short), have become the standard evaluation measures for computational simulations. Following recent studies in the literature we present precision recall and F-scores for boundaries (BP, BR, BF), word tokens (WP, WR, WF) and word types or lexicon (LP, LR, LF). Besides precision and recall, we also present two error measures, oversegmentation (E_o) and undersegmentation (E_u) errors, defined as $E_o = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$ and $E_u = \mathrm{FN}/(\mathrm{FN}+\mathrm{TP})$, where TP, FP, TN and FN are true positives, false positives, true negatives, and false negatives respectively.

In plain words, E_o is the number of the false boundaries inserted by the model divided by the total number of word internal positions in the corpus. Similarly, E_u is the ratio of boundaries missed to the total number of boundaries. Although these error measures are related to precision and recall, they provide different, and sometimes better, insights into the model's behavior.

4.3 Reference models

In this paper, we compare the results obtained by the cue combination model with two baselines. The first baseline is a random model (RM) that assigns boundaries with the probability of boundaries in the input corpus. The RM is more informed than a completely random classifier, but it has been customary (since Brent and Cartwright, 1996) in segmentation literature to set the bar a little bit higher. The second reference model is a lexicon-building model similar to many state-of-the-art models. The model described here, which we call LM, assigns probabilities to possible segmentations as described in Equations 1 and 2.

$$P(s) = \prod_{i=1}^{n} P(w_i) \tag{1}$$

$$P(w) = \begin{cases} (1-\alpha)f(w) & \text{if } w \text{ is known} \\ \alpha \prod_{i=1}^{m} P(a_i) & \text{if } w \text{ is unknown} \end{cases} \tag{2}$$

where s is a sequence of phonemes (e.g., an utterance or a corpus), w_i is the i^{th} word in the sequence, a_i is the i^{th} sound in the word, $f(w)$ is the relative frequency of the word w, m is the number of known words, and $0 \leq \alpha \leq 1$ is the only parameter of the model. In all experiments reported in this paper, we will fix α at 0.5.

For the incremental model defined here, a word is 'known', if it was used in a previous segmentation. The model accepts whole utterances as single words if the utterance does not contain any known words.

model	boundary			word			lexicon		
	P	R	F	P	R	F	P	R	F
Brent (1999)	80.3	84.3	82.3	67.0	69.4	68.2	53.6	51.3	52.4
Venkataraman (2001)	81.7	82.5	82.1	68.1	68.6	68.3	54.5	57.0	55.7
Goldwater et al. (2009)	90.3	80.8	85.2	75.2	69.6	72.3	63.5	55.2	59.1
Blanchard et al. (2010)	81.4	82.5	81.9	65.8	66.4	66.1	57.2	55.4	56.3
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3

Table 1: Performance scores of the reference models LM and RM in comparison with some of the earlier scores reported in the literature. If there were multiple models reported in a study, the result with the highest lexicon F-score is presented. All scores are obtained on the BR corpus.

Table 1 compares the performances of some recent models in the literature using the BR corpus with the two reference models. The LM performs similar to the state-of-the-art models presented in this table. Hence, to aid comparison of the models proposed in this study with the others in the literature, we will (re)report the result of the two baseline models in the rest of this paper. Note that the scores presented in Table 1 can be misleading since the batch models have an advantage due to the way scores are calculated. The scores of the batch models are calculated at the end of training, while scores of the incremental models include initial (presumably bad) choices made before enough exposure to the input. For example, the LM achieves boundary, word and lexicon F-scores of 89%, 81% and 74% respectively, towards the end of the BR corpus. These scores are higher than all of the scores presented in Table 1 (see Table 4 for details the way these scores are calculated).

4.4 Experiments and results

This section reports results of a set of simulations using the modeling framework described so far. All experiments are run on the BR corpus. For all the results reported below, each cue is represented by a set indicators as described in Section 3, multiple indicators for each phoneme n-gram of length one and three are used for left (`l`) and right (`r`) contexts, for all measures that are calculated over phoneme n-grams surrounding the potential boundary. The use of lexical information and lexical stress as standalone strategies are similar to the 'lexicon-building' strategy. The learner inserts complete utterances to the lexicon when the strategy cannot segment the utterance. As the learner starts to learn (from the edges of the sequences in the lexicon) what the edges of words look like, it uses this information to segment later utterances in the input.[3]

We first report the performance results of individual cues, namely, *predictability* (P), *utterance boundaries* (U), *lexical information* (W) and *lexical stress* (S) in Table 2.

Using the predictability cue alone leads to a segmentation performance lower than but close to the state-of-the-art reference model LM. Although these results are not directly comparable to the earlier studies in the literature, the performance scores presented in Table 2 are the best scores presented to date for models using the predictability cue alone. Graphs presented by Brent (1999) indicates about 50%–60% WP and WR and 20%–30% LP for his baseline model utilizing mutual information on the BR corpus. Cohen et al. (2007) report 76% BP, and 75% BR on George Orwell's 1984. Christiansen et al. (1998) report 37% WP and 40% WR with an SRN using phonotactics and utterance boundary cues on another child-directed speech corpus (Korman, 1984).

The model that learns from the utterance boundaries seems to perform the best. The results are comparable, and in some cases better than the LM. Furthermore, the overall scores are also higher than the scores reported by Fleck (2008), where the boundary, word and lexical F-scores were 82.9%, 70.7% and 36.6%, respectively.

Although it is somewhat behind both predictability and utterance boundary cues, the lexical information alone certainly performs better than random. The lower performance of this model in comparison to 'U' suggests that, at least in this setting, phonotactics learned from word tokens found at the utterance edges leads to a better performance compared to the phonotactics learned from the word types in the learner's lexicon.

The experiment that takes only the stress cue into account yields the worst overall results. It seems, when the cue indicates a boundary, it is extremely precise. However, it is also very conservative. This seems to be due to the fact that the model learns to segment at weak–strong transitions, which is expected to be precise. However, since majority of the stress transitions are strong–strong, this covers rather a small portion of the boundaries.

[3]The source code of the application and the data used in this study can be found at `https://bitbucket.org/coltekin/seg/`.

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
P	69.6	92.5	79.5	56.9	70.2	62.9	36.7	49.8	42.3	15.3	7.5
U	82.9	84.8	83.8	70.5	71.7	71.1	33.8	66.9	44.9	6.6	15.2
W	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
S	78.2	8.2	14.8	26.5	9.7	14.2	8.2	38.7	13.5	0.9	92.8
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 2: Results of simulations using individual cues: predictability (P), utterance boundaries (U), lexicon (W) and lexical stress (S). The rows labeled LM and RM are scores of reference models repeated for ease of comparison.

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
PU	82.6	90.7	86.5	72.4	77.4	74.8	42.8	65.3	51.7	7.2	9.3
PUW	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.3	6.7	8.8
PUWS	92.8	75.7	83.4	78.3	68.1	72.9	26.8	62.7	37.5	2.2	24.3
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 3: Results of combination of strategies based on four cues: starting with predictability and utterance boundaries (PU), addition of lexicon (PUW) and lexical stress (PUWS). The rows labeled LM and RM are scores of reference models repeated for ease of comparison.

Table 3 presents combination of predictability and utterance boundaries, followed by lexical information and stress. Here all indicators are combined in a flat, non-hierarchical manner. The combination of predictability and utterance boundaries results in higher F-scores, and it results in more balanced under- and over-segmentation errors. The addition of the lexical information provide a small but consistent improvement. However, adding stress information seems to have an adverse effect. Despite the increased boundary and word precision, all other performance scores go down substantially when we add the stress cue.

The scores in Table 3 are obtained over the complete corpus. As noted in Section 4.3, these scores do not reflect the 'learned' state of the models. Furthermore, we are interested in the progress of a learner as more input is provided. To demonstrate both, E_o and E_u for all combined models are plotted in Figure 1 for each 500 utterances.

An interesting observation that can be made in these graphs is that the models without the stress cue make fewer undersegmentation errors, with the cost of slightly higher oversegmentation. However, the strategy that combines all cues keeps oversegmentation errors low throughout the learning process, and towards the end, it makes fewer undersegmentation errors as well. This suggests that the model combining all cues, including the stress, may be doing better as it collects more evidence. To demonstrate this further, Table 4 presents the same results presented in Table 3, calculated on the last block of an experiment where performance scores were calculated after every 500 input utterances. Besides demonstrating the increase in performance scores when calculated at later stages of learning, the differences between tables 3 and 4 show clearly that despite the fact that it has a detrimental affect when scores are calculated over the complete corpus, the stress cue has a positive effect at the end of the learning process. This suggests that the combined model using stress cue learns slower and makes more mistakes at the beginning. However as evidence accumulates, it starts to be useful, and increases the overall performance of the combined model.

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
PU	85.6	96.7	90.8	78.7	86.0	82.2	71.8	75.9	73.8	6.6	3.3
PUW	83.3	97.2	89.7	75.6	84.5	79.8	69.8	75.5	72.5	7.9	2.8
PUWS	92.5	89.3	90.9	84.2	82.2	83.2	70.9	77.6	74.1	2.9	10.7

Table 4: The same results presented in Table 3, but measured for the last 290-utterances (last block in an incremental experiment with 500-utterance increments).

5 General discussion

This paper introduced an unsupervised and incremental model of segmentation that focuses on combining multiple cues relevant to child language acquisition as attested by earlier studies in psycholinguistics. Unsupervised and incremental models of segmentation that combine multiple cues are not new. There have been many models sharing these properties to some extent. In particular, the model presented in this paper has many similarities with an earlier connectionist model of segmentation presented by Christiansen et al. (1998). However, unlike connectionist models, the model presented here uses accessible explicit representations, and an concrete learning procedure.

Most recent models with explicit representations and statistical learning procedures tend to be models that process their input in 'batch'. These models typically perform better when measured at the overall best performance level, and the insights

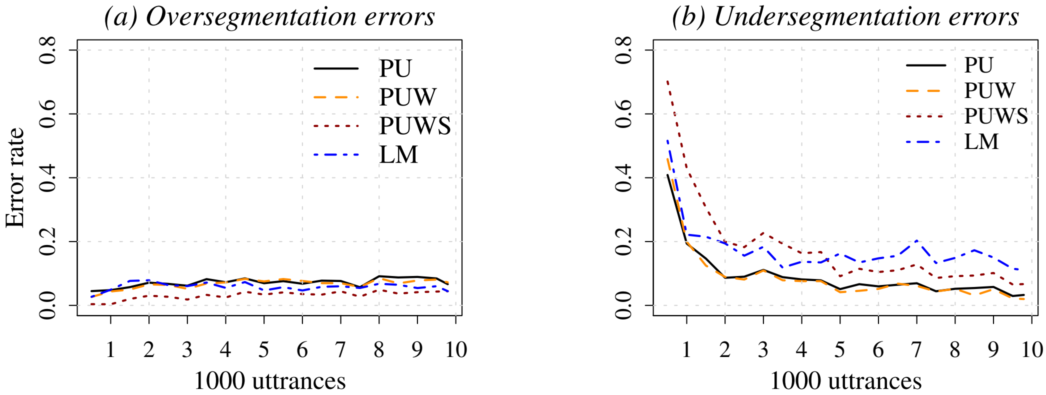

Figure 1: Progression of (a) over- and (b) under-segmentation errors of the combined strategies.

we get from these models are undeniably useful. However, these models typically provide explanations at Marr's (1982) *computational level*. The modeling practice we follow is similar to these models in many ways, and can provide explanations for the same type of questions. However, it may also provide explanations at lower levels (e.g., Marr's *algorithmic level*). This is not to claim that children learn exactly the way the model learns. However, the type of models presented in this paper follow human behavior more faithfully, and, at least in principle, more detailed predictions can be tested on these models. Naturally, relevance of the findings for human cognition will be increased as we constrain our models further in accordance with what we know about the cognitive processes.

The first contribution of this study is the description of a modeling framework that follows what we know about human segmentation process with high fidelity while keeping the benefits of a model with explicit representations and statistical learning methods.

Besides the performance scores that are competitive with the state-of-the-art models in the literature, the simulations also provide some insights regarding the cues commonly studied in the psycholinguistics literature. Some of the findings confirm the previous results. Indeed, it seems that combining multiple cues help. However, the properties of the modeling framework presented in this paper allows us to make some other interesting observations, for example, the effect of stress cue presented in Section 4.4.

When we look at the overall effect of the stress cue throughout the complete simulations, it seems stress degrades the performance. However, if we take a look at the models' performances at the end of the learning, we see that effect of the stress cue is actually positive. In other words, once 'bootstrapped' by the other cues, stress becomes a useful cue. Furthermore, the way the stress cue is useful for the model is also in line with the findings in the literature where stress is commonly found to be a dominant cue (Jusczyk, Cutler, et al., 1993; Thiessen and Saffran, 2003). Given the findings here that stress is rather a precise cue (despite its low recall), it is understandable why it dominates the boundary decisions when available.

The segmentation model presented in this paper demonstrates a way to achieve good segmentation performance using more cognitively relevant and transparent strategies. It is also instrumental at investigating some of the interesting issues regarding cue combination in segmentation, and it is a first step towards models that are more faithful to the human segmentation process. Among other things, we consider two important improvements to the model described here for future work. First, although the combination method used (weighted majority voting) has been successful, other methods such as Bayesian cue combination used for modeling other cognitive processes may be a better approach for segmentation as well. The second improvement we plan is regarding the input. Even though we used a standard corpus as used by many other studies in the literature, it is idealized (e.g., contains no phonetic variation), and poor (e.g., lacking some cues that are available to children) at the same time. Hence, as well as better input representations, using input with variation and noise, and the use of different languages are steps we would like to take in future studies towards a better modeling of segmentation.

References

Richard N. Aslin, Julide Z. Woodward, Nicholas P. LaMendola, and Thomas G. Bever (1996). "Models of Word Segmentation in Fluent Maternal Speech to Infants". In: *Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition*. Ed. by James L. Morgan and Katherine Demuth. Lawrence Erlbaum Associates. Chap. 8, pp. 117–134.

Nan Bernstein Ratner (1987). "The phonology of parent-child speech". In: *Children's language*. Ed. by K. Nelson and A. van Kleeck. Vol. 6. Hillsdale, NJ: Erlbaum, pp. 159–174.

Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer.

Daniel Blanchard, Jeffrey Heinz, and Roberta Golinkoff (2010). "Modeling the contribution of phonotactic cues to the problem of word segmentation". In: *Journal of Child Language* 37.Special Issue 03, pp. 487–511.

Michael R. Brent (1996). "Advances in the computational study of language acquisition". In: *Cognition* 61 (1-2), pp. 1–38.

Michael R. Brent (1999). "An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery". In: *Machine Learning* 34.1-3, pp. 71–105.

Michael R. Brent and Timothy A. Cartwright (1996). "Distributional regularity and phonotactic constraints are useful for segmentation". In: *Cognition* 61 (1-2), pp. 93–125.

Morten H. Christiansen, Joseph Allen, and Mark S. Seidenberg (1998). "Learning to Segment Speech Using Multiple Cues: A Connectionist Model". In: *Language and Cognitive Processes* 13.2, pp. 221–268.

Paul Cohen, Niall Adams, and Brent Heeringa (2007). "Voting experts: An unsupervised algorithm for segmenting sequences". In: *Intelligent Data Analysis* 11.6, pp. 607–625.

Çağrı Çöltekin (2011). "Catching Words in a Stream of Speech: Computational simulations of segmenting transcribed child-directed speech". PhD thesis. University of Groningen.

Anne Cutler and Sally Butterfield (1992). "Rhythmic cues to speech segmentation: Evidence from juncture misperception". In: *Journal of Memory and Language* 31.2, pp. 218–236.

Delphine Dahan and James S. Magnuson (2006). "Spoken Word Recognition". In: *Handbook of Psycholinguistics*. 2nd. Elsevier. Chap. 8, pp. 249–283.

Jeffrey L. Elman (1990). "Finding Structure in Time". In: *Cognitive Science* 14, pp. 179–211.

Margaret M. Fleck (2008). "Lexicalized phonotactic word segmentation". In: *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL-08)*, pp. 130–138.

Timothy Gambell and Charles Yang (2006). *Word segmentation: Quick but not dirty*. Unpublished manuscript.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson (2009). "A Bayesian framework for word segmentation: Exploring the effects of context". In: *Cognition* 112 (1), pp. 21–54.

Katharine Graf Estes, Julia L. Evans, Martha W. Alibali, and Jenny R. Saffran (2007). "Can Infants Map Meaning to Newly Segmented Words? Statistical Segmentation and Word Learning". In: *Psychological Science* 18.3, pp. 254–260.

Zellig S. Harris (1955). "From Phoneme to Morpheme". In: *Language* 31.2, pp. 190–222.

Gaja Jarosz and J. Alex Johnson (2013). "The Richness of Distributional Cues to Word Boundaries in Speech to Young Children". In: *Language Learning and Development* 9.2, pp. 175–210.

Elizabeth K. Johnson and Peter W. Jusczyk (2001). "Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics". In: *Journal of Memory and Language* 44.4, pp. 548–567.

Mark Johnson and Sharon Goldwater (2009). "Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 317–325.

Peter W. Jusczyk (1999). "How infants begin to extract words from speech". In: *Trends in Cognitive Sciences* 3.9, pp. 323–328.

Peter W. Jusczyk, Anne Cutler, and Nancy J. Redanz (1993). "Infants' preference for the predominant stress patterns of English words". In: *Child Development* 64.3, pp. 675–687.

Peter W. Jusczyk, Elizabeth A. Hohne, and Angela Bauman (1999). "Infants' sensitivity to allophonic cues for word segmentation". In: *Perception and Psychophysics* 61.8, pp. 1465–1476.

Peter W. Jusczyk, Derek M. Houston, and Mary Newsome (1999). "The Beginnings of Word Segmentation in English-Learning Infants". In: *Cognitive Psychology* 39, pp. 159–207.

Myron Korman (1984). "Adaptive aspects of maternal vocalizations in differing contexts at ten weeks". In: *First Language* 5, pp. 44–45.

Constantine Lignos (2011). "Modeling infant word segmentation". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 29–38.

Nick Littlestone and Manfred K. Warmuth (1994). "The Weighted Majority Algorithm". In: *Information and Computation* 108.2, pp. 212–261.

Brian MacWhinney and Catherine Snow (1985). "The child language data exchange system". In: *Journal of Child Language* 12.2, pp. 271–269.

David Marr (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Padraic Monaghan and Morten H. Christiansen (2010). "Words in puddles of sound: modelling psycholinguistic effects in speech segmentation". In: *Journal of Child Language* 37.Special Issue 03, pp. 545–564.

Anand Narasimhamurthy (2005). "Theoretical Bounds of Majority Voting Performance for a Binary Classification Problem". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (12), pp. 1988–1995.

Elissa L. Newport and Richard N. Aslin (2004). "Learning at a distance: I. Statistical learning of non-adjacent dependencies". In: *Cognitive Psychology* 48.2, pp. 127–162.

Bruna Pelucchi, Jessica F. Hay, and Jenny R. Saffran (2009). "Learning in reverse: Eight-month-old infants track backward transitional probabilities". In: *Cognition* 113.2, pp. 244–247.

Pierre Perruchet and Stéphane Desaulty (2008). "A role for backward transitional probabilities in word segmentation?" In: *Memory and Cognition* 36.7, pp. 1299–1305.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport (1996). "Statistical learning by 8-month old infants". In: *Science* 274.5294, pp. 1926–1928.

Ivelin Stoianov and John Nerbonne (2000). "Exploring Phonotactics with Simple Recurrent Networks". In: *Proceedings of Computational Linguistics in the Netherlands 1999*. Ed. by Frank van Eynde, Ineke Schuurman, and Ness Schelkens, pp. 51–67.

Kari Suomi, James M. McQueen, and Anne Cutler (1997). "Vowel Harmony and Speech Segmentation in Finnish". In: *Journal of Memory and Language* 36.3, pp. 422–444.

Daniel Swingley (2005). “Statistical clustering and the contents of the infant vocabulary”. In: *Cognitive Psychology* 50.1, pp. 86–132.

Erik D. Thiessen and Jenny R. Saffran (2003). “When Cues Collide: Use of Stress and Statistical Cues to Word Boundaries by 7- to 9-Month-Old Infants,” in: *Developmental Psychology* 39.4, pp. 706–716.

Susan P. Thompson and Elissa L. Newport (2007). “Statistical Learning of Syntax: The Role of Transitional Probability”. In: *Language Learning and Development* 3.1, pp. 1–42.

Anand Venkataraman (2001). “A Statistical Model for Word Discovery in Transcribed Speech”. In: *Computational Linguistics* 27.3, pp. 351–372.