An Incremental Implementation of the Utterance-Boundary Approach to Speech Segmentation

Aris Xanthos

University of Lausanne

Abstract

The problem of speech segmentation is a well-known challenge for various studies, such as language acquisition: how do children correctly infer the position of word boundaries in the continuous stream of speech? One solution to this problem, referred to as the *utterance-boundary strategy*, is to reuse the information provided by the occurrence of specific phonemes sequences at utterance edges in order to hypothesize boundaries inside utterances. In this paper, we describe a probabilistic and incremental implementation of this approach and discuss the results observed for a word segmentation task on a phonemically transcribed and child-oriented French corpus. We show in particular that the first boundaries inferred by this algorithm seem to be reliable enough to make useful generalizations for later decisions.

1 Introduction

The problem of speech segmentation is a well-known challenge for various disciplines, such as linguistics, natural language processing, machine learning and language acquisition studies. In the latter field, it can be informally stated as follows: how do children correctly infer the position of word (or morpheme) boundaries in the continuous stream of speech ?

According to Brent (1999), computational solutions developed so far make use of three general classes of strategies, sometimes combined with one another: a) the *utterance-boundary strategy* consists in reusing the information provided by the occurrence of specific phonemes sequences or prosodic cues at utterance beginnings or endings in order to hypothesize boundaries inside utterances (Aslin, Woodward, Lamendola and Bever 1996, Brent and Cartwright 1996, Christiansen, Allen and Seidenberg 1998); b) the *predictability strategy* assumes that speech should be segmented where the uncertainty about what comes next (phoneme or syllable, for instance) is maximal (Harris 1955, Gammon 1969, Saffran, Newport and Aslin 1996, Hutchens and Adler 1998, Xanthos 2003); c) the *word-recognition strategy* implies an explicit representation of lexical units, and specifies ways of parsing utterances according to the lexicon and of adding novel items to it (Olivier 1968, Wolff 1977, De Marcken 1996, Brent and Cartwright 1996, Brent 1999).

From a cognitive point of view, it is likely that speakers use a "conspiracy" of strategies for speech segmentation (see e.g. Shillcock et al. 2001, Goldsmith 2001), but unless we assume that children have an innate lexicon, and since one-word utterances are not frequent even in child-directed speech[1]), word-recognition

[1] In their corpus of American English, Brent and Siskind (2001) find that "about 9% of infant-direct

can only be performed once other heuristic procedures have sketched a first segmentation. The utterance-boundary strategy is a potential approach to fill that slot, though it has been raised that it could not handle some very frequent words that occur only inside utterances (Brent 1999), and thus might be more appropriate for phrase than for word segmentation.

The utterance-boundary strategy has mainly been implemented using connectionist models[2], unlike the predictability approach, for which many distributional algorithms were designed as well as connectionist ones. In this paper, we describe a probabilistic and incremental implementation of this strategy and discuss the results observed for a word segmentation task on a phonemically transcribed and child-oriented French corpus. We show in particular that the first boundaries inferred by this algorithm seem to be reliable enough to make useful generalizations for later decisions.

2 Algorithm description

2.1 Evaluation of utterance-boundary typicality

In accordance with a psychologically plausible hypothesis (Brent 1999), our algorithm processes the input incrementally, one utterance after another. Intuitively, the idea is to segment utterances where sequences occur, which are typical of utterance boundaries; this should enable us to discover new probable boundary markers, and to improve further segmentation. The specificity of our approach is to formalize this particular typicality in a distributional fashion.

Formally, let $S := \{s_1, \ldots, s_K\}$ be the set of phonemes (or segments) in a language $\mathcal{L}$. $U \subseteq S^*$ is the set of possible utterances in $\mathcal{L}$. $C := u_1 \ldots u_T$, where $u_t \in U$ for $1 \leq t \leq T$, is a corpus of $\mathcal{L}$. For a given order $r \geq 1$, the algorithm works by gradually building three separate distributions[3]:

1. $n(w)$ is the absolute frequency of an r-gram $w \in S^r$ *within* utterances (no overlap *between* utterances);
2. $n(w|I)$ is the absolute frequency of an r-gram in utterance-initial position;
3. $n(w|F)$ is the absolute frequency of an r-gram in utterance-final position.

Relative frequencies may then be defined:

1. $f(w) := n(w)/\sum_{\tilde{w} \in S^r} n(\tilde{w})$ is the relative frequency of an r-gram within utterances;
2. $f(w|I) := n(w|I)/\sum_{\tilde{w} \in S^r} n(\tilde{w}|I)$ is the relative frequency of an r-gram in utterance-initial position;

utterances are isolated words". However, they also demonstrate that the portion of lexicon heard in isolation is rather large.

[2] A significant exception is Brent and Cartwright (1996), where the search space of a word-recognition strategy is restricted only to those sequences that occur at utterance edges. It differs from our approach, however, in that we evaluate *typicality* rather than simply assess *possibility*.

[3] in fact, $3r$ distributions are recorded, see section 2.2 below.

3. $f(w|F) := n(w|F)/\sum_{\tilde{w}\in S^r} n(\tilde{w}|F)$ is the relative frequency of an r-gram in utterance-final position.

Suppose now that we are examining an unsegmented utterance, wondering how much the sequence $w \in S^r$, which occurs there, is typical of utterance endings[4]. Intuitively, the more frequently a sequence occurs in utterance-final position, the more likely it is to be typical of that context, so it seems that our typicality measure should be proportional to $f(w|F)$. But we also need to sort out the case of sequences which are frequent in *any* position; to do this, we can simply divide the frequency of w in utterance-final context by its frequency in any context, thus obtaining the relevant typicality measure[5]:

$$(1) \qquad t(w|F) := \frac{f(w|F)}{f(w)}$$

This measure is higher than 1 iff (if and only if) w is more likely to occur in utterance-final position (than in an unspecified position), lower iff it is less likely to occur there, and equal to 1 iff its probability is independent of its position. For the segmentation procedure, this suggests a “natural” threshold of 1, which can optionally be fine-tuned in order to obtain a more or less conservative result.

2.2 Border effect and symmetry

From a computational point of view, it should be noted that for $r > 1$, there is a *border effect* for boundaries close to utterance edges. For instance, let $u := \alpha_1 \ldots \alpha_m$ ($m \geq 2$) denote an utterance to be processed. Then we cannot compute our typicality measure $t(\alpha_1|F)$ for the first potential boundary (between α_1 and α_2), since we do not know the relevant frequencies for $\tilde{r}$-grams with $\tilde{r} < r$. This implies that we need not only to store the general distribution of r-grams and those in utterance-initial and -final position, but also the corresponding distributions for $1 < \tilde{r} \leq r$, which amount to $3r$ distributions. This will generally remain implicit in the rest of the paper.

Implementations of the utterance-boundary and predictability strategies often combine a “forward” statistics with its “backward” reflection (Harris 1955). Here we do this by simply taking the average of both quantities (namely $t(w|F)$ and $t(w'|I)$, when investigating for a potential boundary between w and $w' \in S^r$). Note that, in cases such as those discussed previously, this can lead to an asymmetric situation when w and w' have different lengths l and l'. In order to compensate for this, we suggest to weight the contributions of the forward and backward measure accordingly:

$$(2) \qquad \bar{t}(w, w') := \frac{l}{l+l'}t(w|F) + \frac{l'}{l+l'}t(w'|I)$$

where $w \in S^l$ and $w' \in S^{l'}$ and $l, l' \leq r$.

[4] We consider only the “forward” case; the derivation of the “backward” case is similar.

[5] Note that taking the log of $t(w|F)$ yields the *pointwise mutual information* (see e.g. Manning and Schütze 1999) between the sequence w and the utterance-final position, which measures the specific dependence between them.

2.3 Parameters updating

As we said in section 2.1, our approach relies on the assumption that, generally, the segmentation of an utterance yields new typical sequences which can be used for further processing. This is implemented by gradually updating the parameters.

At the beginning of the corpus C, i.e. for $t = 0$, the system has no information at all. When a first utterance u_1 is observed, it can get first estimates of $f(\bullet)$, $f(\bullet|I)$ and $f(\bullet|F)$[6]. These are used to segment u_1, and the sequences occurring immediately before and after the newly inferred boundaries are added to the estimates for $f(\bullet|F)$ and $f(\bullet|I)$ respectively[7]. Hopefully, over time, these estimates converge to the "true" (and unknown) distributions in *word*-final and -initial position (denoted by $\hat{f}(\bullet|F)$ and $\hat{f}(\bullet|I)$).

To determine whether or not this convergence is actually observed in a real corpus is the main purpose of the experiments described below. It is expected that using the "true" distributions $\hat{f}(\bullet|F)$ and $\hat{f}(\bullet|I)$ would yield the best possible results of the approach we introduce. Our aim in this paper is not to assess whether this limit is high enough for human or machine language processing, but how close to it an utterance-boundary heuristic might lead.

3 Empirical evaluation

3.1 Experimental setup

The algorithm described in the previous section was implemented and evaluated using a phonemically transcribed and child-oriented French corpus (Kilani-Schoch corpus[8]). For the present research, we have extracted from the original corpus all the utterances of Sophie's parents (mainly her mother) while the child was between ages 1;6.14 and 2;6.25 (year;month.day). This was transcribed phonemically in a semi-automatic fashion, using the BRULEX database (Content, Mousty and Radeau 1990) and making the result closer to oral French with a few handcrafted rules. Eventually the first 10'000 utterances were used for simulations. This corresponds to 37'663 words and 103'325 phonemes (39 types).

In order to evaluate the results of the algorithm (and the effect of various parameters changes) for word segmentation, we compared its output to the segmentation given in the original transcription using traditional measures from the signal detection framework. The *precision* is the probability that an inferred boundary actually occurs in the true segmentation, and the *recall* is the probability for a true boundary to be detected.

Three simulations were run, in an attempt to answer two main questions:

1. How good is the segmentation that our algorithm would perform if it knew

[6] These notations denote the whole distributions of relative frequencies defined in section 2.1.

[7] More complex updating procedures could be designed, for instance by *de*crementing the frequencies of r-grams previously observed and found later to overlap a boundary.

[8] Sophie, a French speaking Swiss child, was recorded at home by her mother every ten days in situations of play (Kilani-Schoch and Dressler 2001). The transcription and coding were done according to CHILDES conventions (MacWhinney 2000).

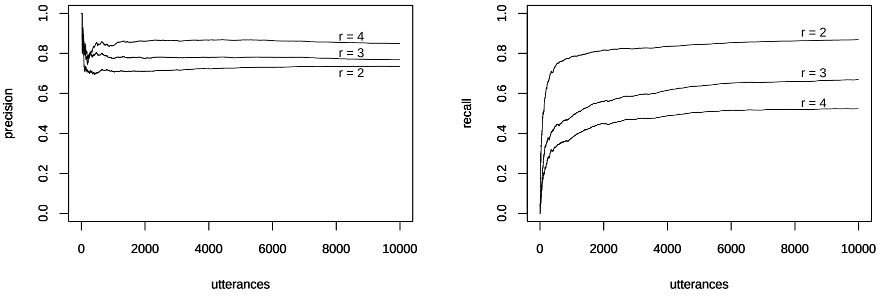

Figure 1: Precision and recall when the "true" distributions in word-initial and -final position are known ($r = 2, 3, 4$).

the "true" distributions in *word*-initial and -final position from the start[9] ?

2. Do the distributions inferred on the basis of sequences in *utterance*-initial and -final position actually converge to the true distributions in word-initial and -final position[10] ?

The results are discussed in the next section.

3.2 Results

3.2.1 First experiment: full information

In this set of simulations, we estimated the distributions of phonemes in *word*-initial and -final position from the whole corpus; these were used as approximations of the corresponding "true" distributions ($\hat{f}(\bullet|I)$ and $\hat{f}(\bullet|F)$). The segmentation was then effected without subsequent updating of the parameters (apart from the r-grams distribution). This should provide the best possible segmentation that we might expect using our algorithm alone.

As can be seen on figure 1, for various orders and a threshold of 1, the precision quickly reaches a stable level, whereas the recall keeps growing with the number of utterances processed (within the limits of our corpus), presumably because the r-grams distribution gets more representative.

Using higher orders amounts to enhance the precision to the prejudice of the recall: concretely, higher orders yield larger phrase-like chunks with "safer" bound-

[9] This amounts to asking how much possible structural properties of speech, such as the high frequency of some words that always occur *inside* phrases (e.g. Eng. *a, the*) can damage the performance of the approach.

[10] Note that this is evaluated here only on the basis of segmentation results; another possibility would be to compute some (dis-)similarity measure over actual distributions.

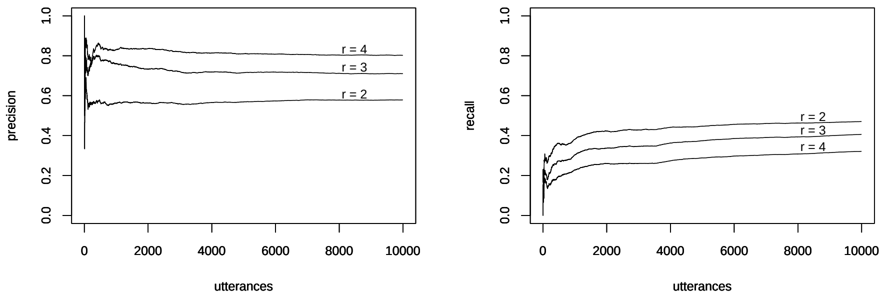

Figure 2: Precision and recall when distributions in utterance-initial and -final position are used, but without generalizing the newly discovered boundaries ($r = 2, 3, 4$).

aries. In general, we can say—at least in this "full information" configuration—that the algorithm favors the precision, which is probably a desirable property, since the ratio of (true) boundaries to (true) non-boundaries in the corpus is (37'663-10'000):(103'325-37'663) = 0.42:1. In other words, for equal precision and recall, we get more than twice as many false alarms as misses[11].

3.2.2 Second experiment: no update

The second set of simulations is intended to show how well the algorithm performs if it does not know the true distributions in *word*-initial and -final position and uses instead the distributions in *utterance*-initial and -final position, but without updating them to discover new typical boundary markers[12].

As shown on figure 2 (see also table 1, p. 177), the performance of this procedure does not compare with that of the previous, supervised one. At the end of the corpus, differences in precision range between .15 for $r = 2$ and .05 for $r = 4$, and differences in recall range between .4 for $r = 2$ and .2 for $r = 4$. In other words, the decrease is more severe for low orders, and for the recall than for the precision. This situation does not change much over time, as the same evolution is observed: stable precision and increasing recall.

3.2.3 Third experiment: update

This experiment demonstrates the normal functioning of the algorithm we propose. As in the "no update" configuration, there is no previous knowledge of the distribu-

[11] Furthermore, undetected boundaries could be detected in a later stage of processing, whereas rectifying false alarms would imply a totally different mechanism.

[12] The distributions are actually updated, but only to reflect the utterance boundaries observed in the input, and not those resulting of the segmentation.

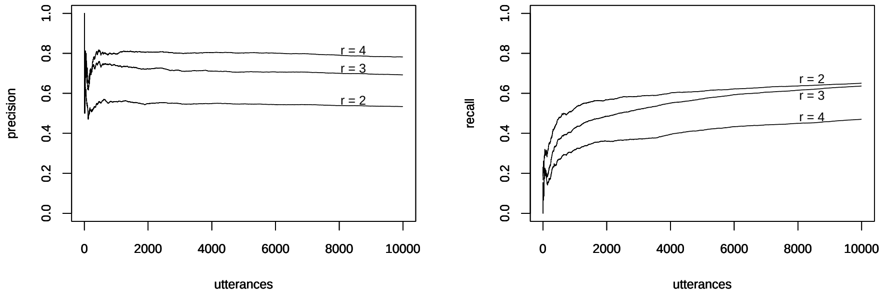

Figure 3: Precision and recall when distributions in utterance-initial and -final position are used and updated to account for the newly discovered boundaries ($r = 2, 3, 4$).

tions in word-initial and -final position, but this time the distributions in utterance-initial and -final position are updated to reflect the newly inferred boundaries. The results are plotted on figure 3 (see also table 1 below).

		precision			recall		
order	mode	3300	6600	10000	3300	6600	10000
2	full info	0.72	0.73	0.73	0.83	0.86	0.87
	no update	0.56	0.58	0.58	0.43	0.46	0.47
	update	0.55	0.54	0.53	0.59	0.63	0.65
3	full info	0.78	0.78	0.77	0.6	0.66	0.67
	no update	0.71	0.72	0.71	0.35	0.39	0.41
	update	0.71	0.71	0.69	0.53	0.6	0.64
4	full info	0.87	0.86	0.85	0.48	0.52	0.52
	no update	0.82	0.81	0.80	0.26	0.3	0.32
	update	0.8	0.8	0.78	0.38	0.44	0.47

Table 1: Summary of the results for the three versions of the algorithm, for $r = 2, 3, 4$ and after approximately 1/3 of the corpus (3300 utterances), 2/3 (6600) and the whole set of utterances.

Though it does not reach the performance of the supervised algorithm, this approach leads to a much better recall than previous procedure, even increasingly better as the corpus size grows. There is of course a slight loss of precision, but it doesn't counterbalance the gain. For instance, with $r = 3$, after processing the whole corpus, we get a recall of .64 for a precision of .69, to be compared with .41 and .71 respectively for the "no update" procedure. Even if we take into account the different frequencies of (true) boundaries and non-boundaries in the corpus, updating the parameters improves the overall accuracy[13] by 2.9%.

It is not surprising that the unsupervised algorithms ("update" as well as "no update") are not as efficient as the "full information" case, for the reason mentioned in section 1: some very frequent words only occur in utterance-internal position, and this is precisely the information that makes the difference between these conditions. However, the better results observed in this last experiment show that generalizing the effected segmentation actually helps recovering part of the missing information—at least in French and for our child-oriented corpus.

4 Conclusions and further issues

In this paper, we have described a probabilistic and incremental implementation of the utterance-boundary strategy for speech segmentation. The method we proposed is unsupervised—in that it does not require an explicit knowledge of the target segmentation—and quite simple as it relies only on (possibly conditioned) r-grams statistics, with no other parameter (recall the "natural" threshold of 1). Yet it gives interesting results in terms of precision and recall, even on a corpus of modest size, though they are clearly too low for a "standalone" segmentation procedure.

We observed that, within this framework, using utterance-boundary typical sequences (second experiment) yields a lower precision and a much lower recall than using "true" word-boundary typical sequences (first experiment)[14]. However, we showed that updating the parameters to take into account the newly inferred boundaries gets much closer to the supervised performance, with a considerable gain in recall for a rather small decrease of precision with respect to the "no update" condition (third experiment).

From the point of view of language acquisition, we believe that the utterance-boundary strategy is well suited as a very first heuristic for segmentation, since it can make correct inferences after processing only a few utterances. Like other strategies, based on assumptions about the metrical structure of the input language (Cutler 1994, Frauenfelder and Content 1999), it relies on perceptually salient features of the data rather than on the less obvious statistical properties used by predictability-based strategies.

However, as witnessed by the recall observed in our experiments, the utterance-boundary strategy has a clear tendency to "under-segment" the data - at least in French. As mentioned earlier, this is due to the fact that some words never occur in utterance-initial or -final position. We have shown that generalizing the results of previous inferences could help making up for this, but still many of the chunks produced by our algorithm are phrases and not words. Thus it seems necessary to hypothesize more strategies in order to get closer to a word-level

[13] defined as the probability for the system to make a correct decision.

[14] By the way, we find it surprising that this quite simple method should give so high results in its supervised version. It suggests that the "typicality" approach can encode efficiently the information brought by a given segmentation; this could be an interesting starting point for another kind of research.

segmentation. The utterance-boundary strategy could then be seen as a tool for “pre-segmentation”—a way of simplifying the data to be processed using other strategies.

Acknowledgments

The author is grateful to Marianne Kilani-Schoch and the mother of Sophie for providing the acquisition corpus (see p. 174), as well as to François Bavaud, Remi Jolivet, Marianne Kilani-Schoch and Christian Lachaud for useful comments on earlier drafts of this paper.

References

Aslin, R., Woodward, J., Lamendola, N. and Bever, T.(1996), Models of word segmentation in fluent maternal speech to infants, *in* J. Morgan and D. K. (eds), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Language Acquisition*, Lawrence Erlbaum Associates, Mahwah (NJ), pp. 117–134.

Brent, M.(1999), Speech segmentation and word discovery: a computational perspective, *Trends in Cognitive Sciences* **3**, 294–301.

Brent, M. and Cartwright, T.(1996), Distributional regularity and phonotactics are useful for segmentation, *Cognition* **61**, 93–125.

Brent, M. and Siskind, J.(2001), The role of exposure to isolated words in early vocabulary development, *Cognition* **81**, 31–44.

Christiansen, M., Allen, J. and Seidenberg, M.(1998), Learning to segment speech using multiple cues, *Language and Cognitive Processes* **13**, 221–268.

Content, A., Mousty, P. and Radeau, M.(1990), Brulex: Une base de données lexicales informatisée pour le français écrit et parlé, *L’Année Psychologique* **90**, 551–566.

Cutler, A.(1994), Segmentation problems, rhythmic solutions, *Lingua* **92**, 81–104.

De Marcken, C.(1996), Linguistic structure as composition and perturbation, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 335–341.

Frauenfelder, U. and Content, A.(1999), The role of the syllable in spoken word recognition: Access or segmentation ?, *Actes des IIèmes Journées d’Etudes Linguistiques*, Université de Nantes, Nantes (France), pp. 1–8.

Gammon, E.(1969), Quantitative approximations to the word, *Papers presented to the International Conference on Computational Linguistics COLING-69*.

Goldsmith, J.(2001), Unsupervised learning of the morphology of a natural language, *Computational Linguistics* **27 (2)**, 153–198.

Harris, Z.(1955), From phoneme to morpheme, *Language* **31**, 190–222.

Hutchens, J. and Adler, M.(1998), Finding structure via compression, *Proceedings of the International Conference on Computational Natural Language Learning*, pp. 79–82.

Kilani-Schoch, M. and Dressler, W.(2001), Filler + infinitive and pre- and protomorphology demarcation in a french acquisition corpus, *Journal of Psycholinguistic Research* **30 (6)**, 653–685.

MacWhinney, B.(2000), *The CHILDES Project: Tools for Analyzing Talk. Third Edition.*, Lawrence Erlbaum Associates, Mahwah (NJ).

Manning, C. and Schütze, H.(1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge (MA).

Olivier, D.(1968), Stochastic grammars and language acquisition mechanisms, Unpublished dissertation, Harvard University.

Saffran, J., Newport, E. and Aslin, R.(1996), Word segmentation: The role of distributional cues, *Journal of Memory and Language* **35**, 606–621.

Shillcock, R., Cairns, P., Chater, N. and Levy, J.(2000), Statistical and connectionist modelling of the development of speech segmentation, *in* P. Broeder and J. Murre (eds), *Models of Language Acquisition: Inductive and Deductive Approaches*, Oxford University Press, Oxford, pp. 103–120.

Wolff, J.(1977), The discovery of segments in natural language, *British Journal of Psychology* **68**, 97–106.

Xanthos, A.(2003), Du k-gramme au mot: variation sur un thème distributionnaliste, *Bulletin de linguistique et des sciences du langage (BIL)*.