Segmenting Speech without a Lexicon: Evidence for a Bootstrapping Model of Lexical Acquisition

Timothy A. Cartwright and Michael R. Brent
Department of Cognitive Science
The Johns Hopkins University
3400 North Charles Street
Baltimore, Maryland 21218
cat@mail.cog.jhu.edu

Abstract

Infants face the difficult problem of segmenting continuous speech into words without the benefit of a fully developed lexicon. Several information sources in speech—prosody, semantic correlations, phonotactics, and so on—might help infants solve this problem. Research to date has focused on determining to which of these information sources infants might be sensitive, but little work has been done to determine the usefulness of each source. The computer simulations reported here are a first attempt to measure the usefulness of distributional and phonotactic information in adult- and child-directed speech. The simulations hypothesize segmentations of speech into words; the best segmentation hypothesis is selected using the Minimum Description Length paradigm. Our results indicate that while there is some useful information in both phoneme distributions and phonotactic rules, the combination of both sources is most useful. Further, this combination of information sources is more useful for segmenting child-directed speech than adult-directed speech. The implications of these results for theories of lexical acquisition are discussed.

Introduction

Infants must learn to recognize phonemic sequences as words; this is a difficult problem because normal speech contains no obvious acoustic cues marking word boundaries. Two sources of information that might aid speech segmentation are: distribution—the phoneme sequence in *cat* appears frequently in several contexts including *thecat*, *cats* and *catnap*, whereas the sequence in *catn* is rare and appears in restricted contexts; and phonotactics—*cat* is an acceptable syllable in English, whereas *pcat* is not. While evidence exists that infants are sensitive to these sources, we know of no measurements of their usefulness. In this paper, we attempt to quantify the usefulness of distribution and phonotactics in segmenting child- and adult-directed speech. Our results show that phonotactics and distribution each provide some information for speech segmentation, but the combination of sources provides substantial information. We also found that child-directed speech is much easier to segment than adult-directed speech when using both sources.

To date, psychologists have focused on two aspects of the speech segmentation problem. The first is the problem of parsing continuous speech into words given a developed lexicon to which incoming sounds can be matched; both psychologists (e.g., Cutler & Carter, 1987; Cutler & Butterfield, 1992) and designers of speech-recognition systems (e.g., Church, 1987) have examined this problem. However, we want to know how infants segment speech before knowing the phonemic forms of words. The second aspect psychologists have focused on is the problem of determining the information sources to which infants are sensitive. The two sources that have been examined most often are: prosody and word stress. Results suggest that parents exaggerate prosody in child-directed speech to highlight important words (Fernald & Mazzie, 1991; Aslin, Woodward, LaMendola & Bever, in press) and that infants are sensitive to prosody (e.g., Hirsh-Pasek et al., 1987). Furthermore, word stress in English fairly accurately predicts the location of word beginnings (Cutler & Norris, 1988; Cutler & Butterfield, 1992), and Jusczyk, Cutler and Redanz (1993) demonstrated that English speaking 9-month-olds (but not 6-month-olds) are sensitive deviations from the predominant strong/weak word stress pattern of English. Sensitivity to native-language phonotactics in 9-month-olds was recently reported by Jusczyk, Friederici, Wessels, Svenkerud and Jusczyk (1993). These studies demonstrate infants' perceptive abilities without demonstrating the usefulness of their perceptions.

In this paper, we measure the potential roles of distribution, phonotactics and their combination using a computer-simulated learning algorithm; the simulation is based on a bootstrapping model in which phonotactic knowledge is used to constrain the segmentation of speech into words.

How do children combine the information they perceive from different sources? Aslin et al. speculate that infants learn words first in isolation, then in context using distribution and prosody to refine their guesses; however, Jusczyk (1993) expresses skepticism about this proposal. One obvious problem with this suggestion is that infants must know when they are hearing single-word utterances—this may be just as difficult as learning to segment multiple-word utterances.

In order to use computer simulations, it is necessary to make certain assumptions about the form of the input. We have tried to make assumptions that are at least consistent with, and often motivated by, research on human infants. For instance, the input to our system is represented as a sequence of phonemes; this was a pragmatic decision and we do not claim that infants convert speech into phoneme sequences during the segmentation process. On the other hand, research

by Kuhl (e.g., Grieser & Kuhl, 1989) suggests that infants may be able to make phonemic distinctions, so this assumption is at least a plausible one. Even if infants do not use phonemic distinctions to segment speech, we believe our system will work given other representations as well—in fact, a narrower (phonetic) transcription system might even help the system (if Church, 1987, is right). Since sentence boundaries provide some information about word boundaries (the end of a sentence is also the end of a word), our input contains sentence boundaries; several studies (Bernstein-Ratner, 1985; Hirsh-Pasek el al., 1987; Kemler Nelson, Hirsh-Pasek, Jusczyk & Wright Cassidy, 1989; Jusczyk et al., 1992) have shown that infants can perceive sentence boundaries using prosodic cues. However, Fisher and Tokura (in press) found no evidence that prosody can accurately predict word boundaries, so the task of finding words remains. Finally, one might question whether infants have the ability we are trying to model—that is, whether they can identify words embedded in sentences; Jusczyk and Aslin (submitted) found that 7 1/2-month-olds can.

It must be emphasized that we did not try to model infant behavior. Our simulations abstract away from the child's computational limitations, such as limited short-term memory, in order to focus on characteristics of the input.

Introduction to the Simulations

The basic results of our experiments come from two variations of a speech segmenting simulation: one which only analyses distributional information and one which analyses distributional information as constrained by phonotactic knowledge.

To gain an intuitive understanding of distributional analysis, consider the following speech sample of three utterances (transcription is in IPA):

Orthography: Do you see the kitty?
See the kitty?
Do you like the kitty?

Transcription: dujusiðəkɪti
siðəkɪti
dujulaɪkðəkɪti

There are many different ways to break this sample into putative words (each particular segmentation is called a segmentation hypothesis). Two such hypotheses are:

Segmentation 1: du ju si ðə kɪti
si ðə kɪti
du ju laɪk ðə kɪti

Segmentation 2: duj us ið əkɪt i
sið ək ɪti
du jul aɪk ðək ɪti

Listing the words used by each segmentation hypothesis yields the following lexicons:

Segmentation 1

1	du	3	kɪti	5	si
2	ðə	4	laɪk	6	ju

Segmentation 2

1	aɪk	5	ək	9	ɪti
2	du	6	əkɪt	10	jul
3	duj	7	i	11	sið
4	ðək	8	ið	12	us

Note that Segmentation 1, the correct hypothesis, yields a compact lexicon of frequent words, whereas Segmentation 2 yields a much larger lexicon of infrequent words. Also note that a lexicon contains only the words used in the sample—no words are known to the system a priori. Given a lexicon, the sample can be represented by replacing words with their lexical indices:

Encoded Sample 1: 1 6 5 2 3
5 2 3
1 6 4 2 3

Encoded Sample 2: 2 12 6 4 5
11 3 8
1 9 10 7 8

The system attempts to find the hypothesis that minimizes the combined sizes of the lexicon and encoded sample. This approach is called the Minimum Description Length (MDL) paradigm and has been used recently in other domains to analyze distributional information (Li & Vitányi, 1993; Rissanen, 1978; Ellison, 1992, 1994; Quinlan & Rivest, 1989; Brent, 1993). The actual system uses more complex representations for the lexicon and the encoded sample that describe them more briefly. For instance, the actual representations assign shorter indices to frequent words than to infrequent words, thus reducing the total number of digits in the encoded sample. For details, see Cartwright and Brent (1994).

The space of possible hypotheses is vast (for our samples, unconstrained by phonotactics, there are about 2.5×10^{406} hypotheses); some method of finding a minimum-length hypothesis without considering all hypotheses is necessary. We used the following method: first, evaluate the input sample with no word boundaries added; then evaluate all hypotheses obtained by adding two word boundaries; take the shortest hypothesis found in the previous step and evaluate all hypotheses obtained by adding two more word boundaries; continue this way until the sample has been segmented into the smallest possible units; finally, report the shortest hypothesis ever found.

In some experimental conditions, phonotactic rules restrict the legal segmentation hypotheses by preventing word boundaries at certain places; for instance, /kætspɔz/ ("cat's paws") has six internal potential word boundaries (k ætspɔz, kæ tspɔz, etc.), only two of which are phonotactically allowed (kæt spɔz and kæts pɔz).

Experiments

We compared results from simulations of two basic experimental conditions: an analysis of distributional information alone (DIST-FREE) and an analysis of distributional information constrained by phonotactic restrictions (DIST-PHONO). Each simulation was run on each of six samples, for a total of twelve DIST runs. Finally, two other simulations were run on each sample to measure chance performance: (1) RAND-FREE randomly inserted word boundaries and reported the resulting hypothesis, (2) RAND-PHONO did the same random insertions subject to the same phonotactic constraints as in DIST-PHONO. Since the RAND simulations were given the number of word boundaries to add (equal to the number of word boundaries needed to produce the natural English segmentation), their performance is an upper bound on chance. In contrast, the DIST simulations must determine the number of word boundaries to add using MDL evaluations. Finally, the reported results for each RAND simulation are the averages computed over 1,000 trials on the same input sample.

Inputs

Two speech samples from each of three subjects were used in the simulations; in one sample a mother was speaking to her daughter and in the other, the same mother was speaking to the researcher. The samples were taken from the CHILDES database (MacWhinney & Snow, 1990) from studies reported in Bernstein (1982). Each sample was checked for consistent word spellings (e.g., 'ts was changed to its), then transcribed into an ASCII-based phonetic representation in a manner that guaranteed that each occurrence of a word was transcribed identically. The transcription system paralleled the IPA alphabet and used one character for each consonant and vowel; dipnthongs, r-colored vowels and syllabic consonants were each represented as one character. For example, "boy" was written as `b7`, "bird" as `bRd` and "label" as `lebL`. Sample lengths were selected to make the number of potential word boundaries nearly equal (about 1,350) when no phonotactic constraints were applied; child-directed samples had 498–536 tokens and 153–166 types, adult-directed samples had 443–484 tokens and 196–205 types. Phonotactic constraints were given to the program as a list of licit beginnings (onsets) and ends (codas) of English syllables; this list was checked against all six samples so that the list was maximally permissive (e.g., the underlined consonant cluster in explore could be divided as ek-splore or eks-plore). Finally, before the samples were fed to the simulations, divisions between words (but not between sentences) were removed.

Results

Results are given in Table 1 below. Each simulation was scored for the number of correct word boundaries inserted, as compared to the natural English segmentation. From the scoring data, two measures of segmentation performance were computed: completeness, the percent of all correct word boundaries that were actually found; and accuracy, the percent of the hypothesized word boundaries that were actually correct. More specifically, completeness was defined as the number of correctly inserted word boundaries (hits or true positives) divided by the number of correct word boundaries (hits plus misses, or true positives plus false negatives); accuracy was defined as the number of correctly inserted word boundaries (hits or true positives) divided by the total number of inserted word boundaries (hits plus false alarms, or true positives plus false positives). Note that there is a trade-off between completeness and accuracy—if all possible word boundaries were added, completeness would be 100% but accuracy would be low; likewise, if only one word boundary was added between two words, accuracy would be 100% but completeness would be low. Accuracy would appear to be more important for children than completeness. For example, deciding 'littlekitty' is a word is less disastrous than deciding 'li', 'tle', 'ki' and 'ty' are all words, because assigning meaning to 'littlekitty' is a reasonable first try at learning word-meaning pairs, whereas trying to assign separate meanings to 'li' and 'tle' is problematic.

One of the most striking implications of Table 1 is that this system segments child-directed speech quite well using phonotactics and distribution (DIST-PHONO). To answer to our

Table 1: Results for all simulations averaged over individual speech samples.

		Simulation			
Target	**Measure**	RAND-FREE	RAND-PHONO	DIST-FREE	DIST-PHONO
Adult	% Completeness	25.1	39.5	95.5	22.5
	% Accuracy	28.9	50.5	36.0	92.0
Child	% Completeness	23.4	40.2	79.9	72.3
	% Accuracy	26.7	51.7	37.4	88.3
Average	% Completeness	24.3	39.9	88.0	46.4
	% Accuracy	27.8	51.1	36.6	89.2

primary question about the relative value of these two sources, we need to compare a number of cells in Table 1. The effect of using phonotactic information can be seen by comparing the average performances of RAND-FREE and RAND-PHONO, which differ only by the addition of phonotactic constraints on segmentations in the latter. Clearly phonotactic constraints are useful, as both completeness and accuracy improve. A similar comparison between the RAND-FREE and DIST-FREE shows that distributional information alone also improves performance. Note in all the results of DIST-FREE that using distributional information alone favors completeness over accuracy; in fact, the segmentation hypotheses produced by DIST-FREE have most words broken into single phoneme units with only a handful of words remaining intact. Two comparisons are needed to show that the combination of distributional and phonotactic information performs better than either source alone: DIST-PHONO compared to RAND-PHONO, for phonotactic information, and to DIST-FREE, for distributional information. The former comparison shows that the sources combined are more useful than phonotactic information alone. The latter comparison is less obvious—the trade-off between completeness and accuracy seems to have reversed, with no clear winner (although the higher accuracy of DIST-PHONO is good). Data on discovered word types helps make this comparison: DIST-FREE found 12% of the words with 30% accuracy and DIST-PHONO found 33% of the words with 50% accuracy. Whereas the word boundary data are inconclusive, word type data demonstrate that combining information sources is more useful than using distributional information alone.

There is no obvious difference in performance between child- and adult-directed speech, except in DIST-PHONO (combined information sources) in which the difference is striking: accuracy remains high and completeness more than triples for child-directed speech. This difference is again supported by word type data: 14% completeness with 30% accuracy for adult-directed speech, 56% completeness with 65% accuracy for child-directed speech.

Discussion

There are three lessons to be drawn from these results: the effectiveness of the combined information sources is an improvement over the effectiveness of either source individually, the combined information sources more effectively segment child-directed speech than adult-directed speech, and the system's overall performance is surprisingly good, considering the limited information used.

The results show a difference between adult- and child-directed speech, in that the latter is easier to segment given both distribution and phonotactics. This lends quantitative support to research which suggests that motherese differs from normal adult speech in ways possibly useful to the language-learning infant (Aslin et al.). In fact, the factors making motherese more learnable might be elucidated using this technique: compare the results of several different models, each containing a different factor or combination of factors, looking for those in which a substantial performance difference exists between child- and adult-directed speech.

Our technique segments continuous speech into words using only distributional and phonotactic information more effectively than one might expect—up to 66% completeness of word boundaries with 92% accuracy on one sample, which yields 58% completeness of word types with 67% accuracy (the relatively low type accuracy is mitigated by the fact that most incorrect words are meaningful concatenations of correct words—e.g., 'thekitty'). This finding confirms the idea that distribution and phonotactics are useful sources of information that infants might use in discovering words (e.g., Jusczyk et al., 1993b). In fact, it helps explain infants' ability to learn words from parental speech: these two sources alone are useful and infants have several others, like prosody and word stress patterns, available as well. It also suggests that semantics and isolated words need not play as central a role as one might think (e.g., Jusczyk, 1993, downplayed the utility of words in isolation). It is difficult, if not impossible given currently available methods, to determine which sources of information are necessary for infants to segment speech and learn words; only this sort of indirect evidence is available to us.

Our model uses phonotactic constraints as absolute requirements on the structure of individual words; this implies that phonotactics have been learned prior to attempts at segmentation. We must therefore show that phonotactics can indeed be learned without access to a lexicon—without such a demonstration, we are trapped in circular reasoning. Gafos and Brent (1994) suggest that this may be true. This does not mean we believe infants necessarily learn their native language's phonotactic rules prior to learning to segment speech; rather, we think it is important to investigate how far such an assumption would go toward explaining children's early acquisition of word sounds.

Conclusions and Future Work

Until now, research has focused on demonstrations of infants' sensitivity to various information sources in the input. We have now begun to supplement this evidence with quantitative measures of the usefulness of those sources.

We plan to extend the computer system described here by adding syllable stress (strong or weak) to the representation of words. With this elaborated representation, the MDL paradigm can be used to learn and make use of the predominant stress patterns in the language. We expect that this work will shed light on the utility of stress patterns in combination with the information sources investigated in this paper. Further, if the modified system is able to learn the predominant stress pattern, that will help to resolve a problem with current theories about the use of stress in lexical acquisition; namely, that stress is a language-particular property, and hence must be learned before it can be used.

References

Aslin, R. N., Woodward, J. Z., LaMendola, N. P. & Bever, T. G. (in press). Models of word segmentation in fluent maternal speech to infants. In Morgan & Demuth (Eds.), *Signal to Syntax: Bootstrapping from Speech to Syntax in Early Acquisition*. Hillsdale, NJ: Erlbaum.

Bernstein, N. (1982). Acoustic study of mothers' speech to language-learning children: An analysis of vowel articulatory characteristics. Unpublished doctoral dissertation, Boston University.

Bernstein-Ratner, N. (1985, November). Cues which mark clause-boundaries in mother-child speech. Paper presented at the meeting of the American Speech-Language Hearing Association, Washington DC.

Brent, M. (1993). Minimal generative explanations: A middle ground between neurons and triggers. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 28–36). Hillsdale, NJ: Erlbaum.

Cartwright, T. A. & Brent, M. R. (1994). Segmenting speech without a lexicon: The roles of phonotactics and speech source. In S. Bird (Ed.), *Proceedings of the First Meeting of the ACL Special Interest Group in Computational Phonology* (pp. 83–90). Association for Computational Linguistics.

Church, K. (1987). Phonological parsing and lexical retrieval. *Cognition*, 25, 53–69.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory & Language*, 31, 218–236.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133–142.

Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *JEP: Humen Perception and Performance*, 14, 113–121.

Ellison, T. M. (1992). The Machine Learning of Phonological Structure. Unpublished doctoral dissertation, University of Western Australia.

Ellison, T. M. (in press). The iterative learning of phonological rules. *Computational Linguistics*.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, 27, 209–221.

Fisher, C., & Tokura, H. (in press). Acoustic cues to clause boundaries in speech to infants: Cross-linguistic evidence. In Morgan & Demuth (Eds.), *Signal to Syntax: Bootstrapping from Speech to Syntax in Early Acquisition*. Hillsdale, NJ: Erlbaum.

Gafos, A., & Brent, M. R. (1994). Learning syllable structure without word boundaries. Paper presented at the 1994 Stanford Child Language Research Forum, Stanford, CA.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Wright Cassidy, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, 26, 269–286.

Grieser, D., & Kuhl, P. K. (1989). The categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, 25, 577–588.

Jusczyk, P. W. (1993). Discovering sound patterns in the native language. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 49–60). Hillsdale, NJ: Erlbaum.

Jusczyk, P. W., & Aslin, R. N. (submitted for publication). Recognition of familiar patterns in fluent speech by 7 1/2-month-old infants.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, 64, 675–687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory & Language*, 32, 402–420.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, 24, 252-293.

Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Wright Cassidy, K. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, 16, 55–68.

Li, M., & Vitányi, P. (1993). *An Introduction to Kolmogorov Complexity and its Applications*. New York, NY: Springer-Verlag.

MacWhinney, B., & Snow, C. (1990). The Child Language Data Exchange System: An update. *Journal of Child Language*, 17, 457–472.

Quinlan, J. R., & Rivest, R. L. (1989). Inferring decision trees using the minimum description length principle. *Information and Computing*, 80, 227–248.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465–471.