The Use of Domain-initial Strengthening in Segmentation of Continuous English Speech

James M. McQueen and Taehong Cho
Max Planck Institute for Psycholinguistics
Postbus 310, 6500 AH Nijmegen, The Netherlands
E-mail: james.mcqueen@mpi.nl, taehong.cho@mpi.nl

ABSTRACT

Prosodic structure in English speech is signalled, in part, by stronger articulation of consonants at the onset of intonational phrases (IPs) than of consonants that are IP-medial. In two cross-modal priming experiments, American English listeners heard sentences and decided whether visual letter strings, presented during the sentences, were real words. We manipulated sentence type (either no IP boundary or an IP boundary in a critical two-word sequence), splicing (whether the onset of the sequence's second word was spliced from another token of that sentence or cross-spliced from a matched sentence with or without an IP boundary), and relatedness (whether the visual target was the first word in the spoken sequence). There was a relatedness effect on target responses for sentences with no IP boundary only when they were cross-spliced, that is, where splicing provided evidence of domain-initial strengthening. Listeners thus use this evidence when segmenting continuous speech.

1. INTRODUCTION

The prosodic structure of English speech is marked, at least in part, by the relative strength of segments in domain-initial position [1,2,3]. Such segments are more strongly articulated in larger prosodic domains. For example, consonants in initial position in an intonational phrase (IP) are likely to be longer than word-initial consonants embedded within a phrase. In this paper, we ask if the acoustic-phonetic fine detail associated with IP-initial consonants is used by listeners to segment continuous speech into words. That is, we examine not only whether listeners are sensitive to domain-initial strengthening, but also how they might use such information in continuous speech recognition.

All current models of human spoken-word recognition assume that, as a listener hears an utterance, the words that are consistent with different portions of the input are activated, and that these multiple candidate words compete with each other (see [4] for a review). This process results in the segmentation of continuous speech into words: As candidate words win the competition, word boundaries are "found" between them. This competition process, however, is also modulated by cues that are present in the speech input which signal likely word boundaries [5]. Metrical, allophonic and phonotactic information all appear to influence lexical segmentation in this way. Furthermore, fine-grained phonetic detail in the signal, such as the duration of individual segments, modulates lexical segmentation, as shown across a range of languages and experimental tasks [6,7,8,9].

2. EXPERIMENT 1

We used the cross-modal identity priming task to test if the acoustic manifestation of domain-initial strengthening influences segmentation. American English listeners heard American English sentences as they saw letter strings on a computer screen. Their task was to listen to the sentences, and to decide whether the letter strings were real words.

Each experimental sentence contained a critical two-word sequence, separated by a prosodic word boundary. These two-word sequences were partially lexically ambiguous. The first word plus the onset of the second word was always consistent with at least one other English word. This lexical competitor was intended to make it harder for listeners to segment the two-word sequences correctly.

The effect of domain-initial strengthening was tested by cross-splicing the sentences. In one version of each sentence (the identity-spliced version), the first CV of the second word in the sequence was spliced from another token of that sentence. But in a second version of the sentence (the cross-spliced version) the onset of the second word was spliced from a matched sentence in which an IP boundary occurred within the same two-word sequence. If domain-initial strengthening can be used in segmentation, recognition of the first word should be easier in the cross-spliced sentences than in the identity-spliced sentences. Note that since we were interested in the segmentation process, we manipulated the degree of domain-initial strengthening in the second word, but measured recognition of the first word.

Ease of segmentation of the sequence was measured by manipulating the relationship between the first word and the visual target. Faster lexical decision responses to a target when it is the same as the first word in the sequence than when it is unrelated to the spoken material can be taken to reflect substantial activation of the first word in the

 ISBN 1-876346-48-5

speech recognition system. The question, therefore, was whether this priming effect would be stronger in the cross-spliced than in the identity-spliced sentences.

2.1. METHOD

2.1.1. Participants. Forty seven volunteers from Ohio State University (OSU) were paid for their participation. They were all speakers of American English, with no known hearing problems.

2.1.2. Materials. Forty eight pairs of English sentences were constructed. The same critical two-word sequence appeared in both sentences within a pair, for example, the sequence *bus tickets* in the pair:

(1) John forgot to buy bus # tickets for his family (# = Wd)

(2) When you get on the bus, # tickets should be shown to the driver (# = IP)

In each two-word sequence, as mentioned earlier, the first word plus the onset of the second word was another word or the beginning of another word (e.g., *bust* in *bus tickets* or *partner* in *part names*). The initial consonant of the second word was either a voiceless stop ([p], [t], or [k]), the fricative [s] or the nasal [n]. In one sentence in each pair, there was a prosodic word boundary (Wd) between the two critical words. In the other sentence there was an IP boundary at that location. Sentences of type (1) were the experimental sentences. Those of type (2) were used for cross splicing. The 48 first words of each two-word sequence served as visual targets (e.g., *bus*).

A further 112 filler and 10 practice sentences were constructed. Each of these sentences contained a two-word sequence with a Wd boundary. Forty eight of these sequences were used to make matched sentences with an IP boundary between the two words; these IP sentences were used to make cross-spliced filler sentences. For 24 of the spliced filler sentences, nonword targets were made that were phonologically related to the first word in the critical pair (e.g., for the sequence *fine diamonds*, the nonword *fipe*, which has the same onset and vowel as *fine*). For the other 24 spliced filler sentences, and another 56 unspliced fillers, phonologically unrelated nonword targets were made (e.g., for the sequence *dam project*, the nonword *frist*). Finally, real words that were unrelated to the first word in each designated pair were selected as targets for the remaining 32 filler sentences (e.g., for the sequence *special glue*, the word *hamster*).

2.1.3. Procedure. Multiple tokens of each sentence were recorded in a sound-damped booth by a male native speaker of American English. Two versions (identity vs. spliced) of each experimental sentence of type (1) were then made using the Praat speech editor. The same carrier token of each sentence was used in each version. The identity-spliced version was made by splicing the initial CV of the post-boundary word (e.g., the [tɪ] of *tickets*) from another token of that sentence into the carrier sentence. The cross-spliced version was made by splicing into the same carrier the initial CV of the post-boundary word from a token of the matched IP boundary sentence. The onset of each CV was defined as (1) the release of the closure (for stops), (2) the beginning of the high-frequency frication noise (for [s]) or (3) the beginning of the nasal murmur (for [n]). All splices were made at zero-crossings. Further, the two versions of each sentence were equated for F0 differences using the PSOLA resynthesis method, based on mean F0 values of the paired spliced CVs. The same splicing procedure was applied to the 48 filler sentences with matched IP sentences (24 were identity-spliced and 24 were cross-spliced). The other filler and practice sentences were not spliced.

Each listener heard all the sentences once, and saw all of the targets once. Splicing (identity-spliced vs. cross-spliced) and relatedness (visual target identical to the first word in the critical sequence or unrelated to it) were counterbalanced across four lists. Thus, for example, across the four lists, the visual target *bus* was paired with the identity- and cross-spliced versions of the spoken *bus tickets* sentence, and with the two spliced versions of an unrelated sentence (that with the sequence *mill company*). Each list contained all the filler trials. Any given participant saw 48 experimental and 32 filler word targets and 80 nonword targets. Twenty four of the words (i.e., 15% overall) were related to the sentence being heard (i.e., were identical to the first word in the critical two-word sequence); 24 nonwords were phonologically related to the first word in these sequences. Furthermore, each type of spliced sentence was just as likely to be paired with a word as with a nonword target.

Participants were tested individually in a quiet room. The sentences were presented over headphones, and the targets appeared in lower case on the screen of a laptop computer. The targets appeared on the screen aligned in time with the offset of the first word in each critical sequence, and remained on the screen for 1 second. Participants were asked to listen to the sentences, and to decide as quickly and as accurately as possible whether the target letter strings were real English words. They had to respond by pressing one of two buttons, labelled "YES" and "NO". All participants received the practice trials, followed by one of the four lists. Participants were also informed that they would be given a comprehension test at the end of the experiment. This test comprised 16 written sentences; half of them had been presented auditorily, and half were new. Participants had to judge whether these sentences had appeared in the main part of the experiment. They made, on average, 11.9 (74%) correct responses to these sentences.

2.2. RESULTS AND DISCUSSION

Lexical decision reaction times (RTs) were measured from onset of the visual presentation of the target words. Responses slower than 1200 ms were treated as errors (less than 1% of the data). Responses to two targets with error rates greater than 10% were excluded from the analysis. Mean RTs are shown in Figure 1.

ISBN 1-876346-48-5

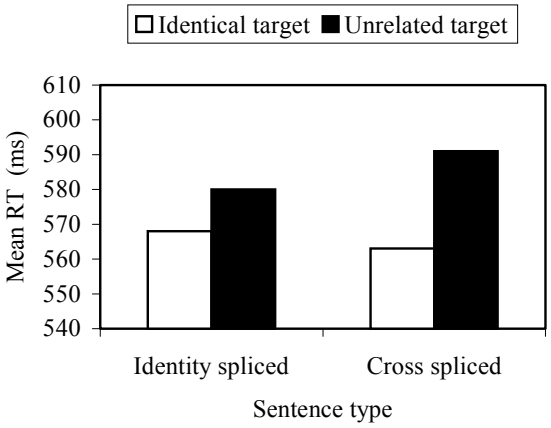

Figure 1. Mean lexical decision latencies, Experiment 1.

Analyses of Variance (ANOVAs) on the RT data treated participants (*F1*) or items (*F2*) as the repeated measure. They revealed an identity priming effect: Listeners were faster to respond to identical targets (e.g., to *bus* as *"bus tickets"* was heard) than to unrelated targets (e.g., to *bus* as *"mill company"* was heard): $F1(1,43) = 16.6$, $p < .001$; $F2(1,45) = 5.6$, $p < .05$. There was no main effect of splicing, but there was a trend towards an interaction of this factor with the priming effect: $F1(1,43) = 3.1$, $p = .08$; $F2(1,45) = 1.5$, $p > .1$. Planned pairwise comparisons showed that the priming effect was significant for the cross-spliced sentences (i.e., where the onset of the post-boundary word was spliced from IP-initial position; $t1(46) = 3.5$, $p < .005$; $t2(45) = 2.6$, $p < .05$) but not for the identity-spliced sentences ($t1(46) = 2.0$, $p = .05$; $t2(45) = 1.1$, *ns*). Further analyses suggested that this interaction was largely due to the items in which the splicing involved stops. There were no significant effects in ANOVAs on error rates. The overall mean error rate was 2%.

Listeners were therefore faster to recognise a target word such as *bus* when they were hearing a sentence containing *bus tickets* than when they were hearing an unrelated sentence, but only when the onset of *tickets* originated from IP-initial position. We suggest that the acoustic-phonetic cues associated with domain-initial strengthening (particularly those for stops) helped listeners to segment the sequences (i.e., helped them rule out competitor words, such as *bust* in *bus tickets*). The resulting strong activation of the correct first word in the sequence led to the priming effect seen in visual lexical decision. In contrast, when no strengthening cues were available (i.e., in the identity-spliced sentences), given the added difficulties due to the presence of a lexical competitor, listeners tended to be unable to complete their segmentation of the critical sequence by the time they were initiating their lexical decision responses, so no priming was observed.

3. EXPERIMENT 2

In our second experiment, we asked whether domain-initial strengthening could also be of benefit to lexical segmentation in the context of an IP boundary. We therefore ran a direct analogue of Experiment 1, using the same task and experimental manipulations, and indeed the same critical materials, but using as the carrier sentences those with IP boundaries that were used for cross-splicing in Experiment 1 (the type (2) sentences). Thus listeners heard sentences including, for example, *"... bus, tickets ...",* in which the initial CV of the post-boundary word came either from another token of that IP boundary sentence and thus maintained strengthening cues, or from the matched sentence with a Wd boundary (the type (1) sentences) and thus did not contain strengthening cues.

If domain-initial strengthening facilitates lexical segmentation in this context, there should be a stronger priming effect for identity-spliced than for cross-spliced sentences (note that this is opposite to the pattern observed in Experiment 1). It was possible, however, that sufficient segmentation cues would be available in both types of sentence given the presence of an IP boundary. Pre-boundary lengthening and boundary tones, for example, are robust phonetic correlates of prosodic structure [10,11,12,13] and could be used in lexical segmentation. Since this information is present in both the identity- and cross-spliced sentences, segmentation (and hence the priming effect) could be equivalent across conditions.

3.1. METHOD

3.1.1. Participants. Forty eight new volunteers from OSU were paid to take part. They were all speakers of American English, with no known hearing problems.

3.1.2. Materials. The experimental sentences were the 48 sentences with IP boundaries that were used for cross-splicing in Experiment 1. There were again 48 filler sentences that were also spliced. These were the sentences with IP boundaries that were matched to the equivalent spliced fillers in Experiment 1. The other filler sentences (64 fillers and 10 for practice) were new; all of these had an IP boundary near the middle of the sentence. The visual targets (words and nonwords) were the same as before.

3.1.3. Procedure. The new sentences were recorded by the same speaker, during the Experiment 1 recording session. The same procedure as in the first experiment was used for speech editing, counterbalancing materials across lists, and running the experiment. The only difference between the experiments, therefore, was that the visual targets now appeared during two-word sequences that contained an IP boundary. Listeners made, on average, 11.6 (73%) correct responses on the comprehension test.

3.2. RESULTS AND DISCUSSION

Lexical decision RTs (from target onset) slower than 1200 ms were again treated as errors (less than 1% of the data). Responses to one target (error rate > 10%) were excluded from the analysis. Mean RTs are shown in Figure 2.

ANOVAs on these data showed that listeners were again faster to respond to identical targets (e.g., to *bus* as *"bus,*

ISBN 1-876346-48-5

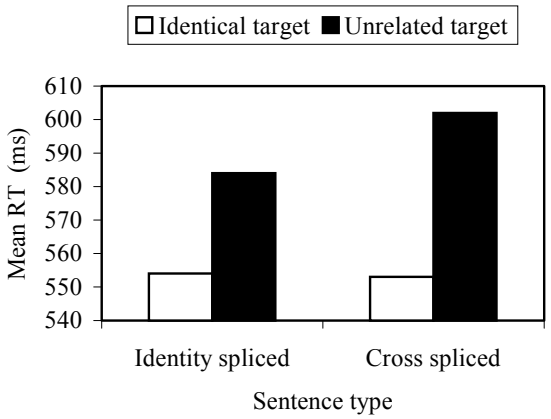

Figure 2. Mean lexical decision latencies, Experiment 2.

tickets" was heard) than to unrelated targets (e.g., to *bus* as *"mill, company"* was heard): $F1(1,44) = 57.5$, $p < .001$; $F2(1,46) = 14.7$, $p < .001$. There was no main effect of splicing, but again a trend towards an interaction of splicing with priming: $F1(1,44) = 3.3$, $p = .08$; $F2(1,46) = 3.9$, $p = .05$. Pairwise comparisons showed, however, that the priming effect was significant for both the identity-spliced sentences ($t1(47) = 5.3$, $p < .001$; $t2(46) = 2.7$, $p < .001$) and the cross-spliced sentences ($t1(47) = 6.1$, $p < .001$; $t2(46) = 4.1$, $p < .001$. The overall mean error rate was again 2%, and there were again no significant effects in the error analyses.

In this experiment, therefore, there was no evidence that the acoustic cues to domain-initial strengthening in the second word in a two-word sequence that spanned an IP boundary influenced recognition of the first word. It would appear that the domain-final cues in the first word (i.e., domain-final lengthening and/or a boundary tone on *bus* in *bus, tickets*) were sufficiently powerful to allow listeners to segment the sequence correctly (and thus rapidly rule out spurious candidate words such as *bust*), irrespective of the nature of the onset of the second word.

4. CONCLUSIONS

These results suggest that listeners are sensitive to domain-initial strengthening. More specifically, they suggest that the phonetic fine detail associated with initial strengthening can be used in lexical segmentation, such that the strengthening of the initial portion of the second word in two-word sequences assists in segmentation of that sequence (particularly when that portion contains a stop; Experiment 1). It appears, however, that when other boundary cues are available (e.g., pre-boundary lengthening, Experiment 2), domain-initial strengthening plays a lesser role in segmentation.

ACKNOWLEDGEMENTS

We thank Mark Pitt for making testing facilities available to us at OSU.

REFERENCES

[1] C. Fougeron and P.A. Keating, "Articulatory strengthening at edges of prosodic domains," *Journal of the Acoustical Society of America*, vol. 101, pp. 3728–3740, 1997.

[2] J. Pierrehumbert and D. Talkin, "Lenition of /h/ and glottal stop," in *Papers in Laboratory Phonology II*, G.J. Docherty and D.R. Ladd, Eds., pp. 90–117. Cambridge, Cambridge University Press, 1992

[3] T. Cho. *The Effects of Prosody on Articulation in English*. New York NY: Routledge, 2002.

[4] J.M. McQueen, "Speech perception," in *The Handbook of Cognition*, K. Lamberts and R. Goldstone, Eds. London: Sage Publications, in press.

[5] D. Norris, J.M. McQueen, A. Cutler and S. Butterfield, "The possible-word constraint in the segmentation of continuous speech," *Cognitive Psychology*, vol. 34, pp. 191–243, 1997.

[6] D.W. Gow and P.C. Gordon, "Lexical and prelexical influences on word segmentation: Evidence from priming," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, pp. 344–359, 1995.

[7] P. Tabossi, S. Collina, M. Mazzetti and M. Zoppello, "Syllables in the processing of spoken Italian," *Journal of Experimental Psychology: Human Per-ception and Performance*, vol. 26, pp. 758–775, 2000.

[8] M.H. Davis, W.D. Marslen-Wilson and M.G. Gaskell, "Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 28, pp. 218–244, 2002.

[9] E. Spinelli, J.M. McQueen and A. Cutler, "Processing resyllabified words in French," *Journal of Memory and Language*, vol. 48, 233–254, 2003.

[10] I. Lehiste, "The timing of utterances and linguistic boundaries," *Journal of the Acoustical Society of America*, vol. 51, pp. 2018–2024, 1972.

[11] M.E. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English" *Phonology Yearbook*, vol. 3, pp. 255–309, 1986.

[12] M.E. Beckman and J. Edwards, "Lengthenings and shortenings and the nature of prosodic constituency," in *Papers in Laboratory Phonology I*, J. Kingston and M.E. Beckman, Eds., pp. 152–178. Cambridge, Cambridge University Press, 1990.

[13] C. Wightman, S. Shattuck-Hufnagel, M., Ostendorf and P. Price, "Segmental durations in the vicinity of prosodic phrase boundaries" *Journal of the Acoustical Society of America*, vol. 91, pp. 1707–1717, 1992.