in *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*
Erlbaum: Hillsdale NJ
Final version, April 1997

Ambiguity and Competition in Lexical Segmentation

Matt H. Davis, William D. Marslen-Wilson and **M. Gareth Gaskell**
Centre for Speech and Language
Psychology Department
Birkbeck College
Malet Street
London WC1E 7HX.
{m.davis,w.marslen-wilson,g.gaskell}@psyc.bbk.ac.uk

Abstract

Earlier research has suggested that left embedded words (e.g. *cat* in *catalog*) present a problem for spoken word recognition since it is potentially unclear whether there is a word boundary at the offset of *cat*. Models of spoken word recognition have incorporated processes of competition so that the identification of embedded words can be delayed until longer interpretations have been ruled out. However, evidence from acoustic phonetics has previously shown that there are differences in acoustic duration between the syllables of embedded words and the onsets of longer competitors. The research reported here used gating and cross-modal priming to investigate the recognition of embedded words. Results indicate that subjects use these acoustic differences to discriminate between monosyllabic words and the onset of longer words. We therefore suggest that on-line processes of lexical segmentation and word recognition are sensitive to acoustic information, such as syllable duration, that may only be contrastive with reference to prior spoken context.

Introduction

This paper is concerned with the recognition of words in connected speech. A substantial part of this problem for both human and machine recognition is the identification of boundaries between words. Connected speech is continuous, there are no gaps between words equivalent to those in written text (Lehiste, 1972), nor are there acoustic cues that reliably mark the position of boundaries between words (Nakatani and Dukes, 1977).

Accounts of how the speech stream comes to be segmented into lexical items[1] can be divided into two main classes. The first type of account describes ‘pre-lexical’ segmentation strategies based on cues that can be used to identify boundaries prior to, or in the absence of, lexical access to words in the speech stream. Examples of pre-lexical cues that have been suggested in the literature include metrical stress (Cutler and Norris, 1988), phonotactic information provided by transitional probabilities (Hayes and Clark, 1970; Cairns et al., 1994) and prosodic boundaries (Nakatani and Schaffer, 1978)

However, these cues do not reliably mark all word boundaries and therefore cannot be relied upon to segment speech. Consequently, the second class of account of segmentation focuses on the role of the recognition process in dividing the speech stream in to words. For example, in models like TRACE (McClelland and Elman, 1986), competition between lexical hypotheses that span potential word boundaries ensures that only words making up a consistent segmentation of the speech stream are activated.

Recognizing Embedded Words

One case in which lexical competition has been argued to be especially valuable is in the identification of words that are embedded at the onset of longer words. These left-embedded words (e.g. *cat* embedded in *catalogue*) may present a particular problem to models of spoken word recognition since at the offset of a syllable like *cat* it is unclear whether what has been heard is a short word or the start of a longer word.

Corpus searches have shown that these onset-embeddings are common in English. Luce (1986) showed that 41% of words in a 20 000 word dictionary are non-unique at their offset, similarly McQueen et al. (1995) have shown that 58% of polysyllabic words contain a shorter word embedded at their onset. Consequently it has been suggested that models of spoken word recognition must provide an account of the recognition of embedded words.

In models such as TRACE, it is competition between lexical hypotheses that allows the resolution of this potential ambiguity. Mutually inhibitory connections between lexical units spanning word boundaries allows the identification of embedded words to be delayed until following context can be used to rule out longer interpretations. Similar results have been obtained in recurrent networks trained to preserve lexical activation until following context becomes available (Content and Sternon, 1994). The use of following context to identify embedded words will delay recognition until after their acoustic offset, as has been confirmed in gating experiments (Grosjean, 1985; Bard et al., 1988).

[1] Accounts of segmentation have almost exclusively assumed that the unit of lexical representation is a dictionary word. For an alternative suggestion see Marslen-Wilson (1996).

Acoustic Phonetics

The emphasis on the use of following context and delayed recognition is based on an assumption that there is ambiguity between tokens of words like *cat* and the first syllable of longer words like *catalogue*. This is despite evidence from acoustic phonetics suggesting that there are consistent differences between the acoustic realization of syllables in monosyllabic and bisyllabic words.

Lehiste (1972), for example, reports significant differences in the acoustic duration of the syllable [slIp] in words such as *sleep*, *sleepy* and *sleepiness*. This was further quantified by Klatt (1976), showing that syllables were 15% shorter in polysyllabic words than when the equivalent syllable was produced as a monosyllable[2].

The goal of the current research was to take a fresh look at the recognition of embedded words such as *cat* in *catalogue*. In particular we were investigating listeners sensitivity to the acoustic cues that differentiate between syllables in short and long words. Models of spoken word recognition such as TRACE don't provide any mechanism for using these acoustic cues, and would therefore predict no differences between the processing of embedded words and phonetically identical onsets of longer competitors. Consequently any differences in the processing of syllables from short and long words would present a challenge to current models.

We were also looking for evidence of competition between short and long word hypotheses. Either such that the recognition of embedded words is delayed by activation of longer competitors or such that short words are activated during the identification of these longer words. In investigating these issues we used two different methods for tapping into the activation of words in the speech stream - gating and cross-modal priming of lexical decision.

Experimental Stimuli

Since we were interested in tapping into lexical level processes in the segmentation of connected speech, experimental stimuli were created with the intention of creating the maximum amount of ambiguity in the position of word boundaries. An automated search of the CELEX database (Baayen et al., 1995) was therefore carried out to find an appropriate set of embedded words.

Forty pairs of words such as *cap* and *captain* were selected. All of the pairs were monosyllabic words embedded at the onset of stress-initial bisyllables, with the syllable boundary of the longer word being at the offset of the embedded word (i.e. the syllable boundary in *captain* is at the offset of *cap*). The short and long words were morphologically unrelated as well as being matched for syntactic class and frequency of occurrence[3].

In order to investigate the recognition of these words in connected speech, non-biasing sentence contexts were generated for each pair of words. A multiple-cloze test was carried out on these contexts, to ensure that none of the test words were predictable from the context. An 'ease-of-completion' rating task was also carried out to ensure that the target words didn't differ in the ease with which they could be interpreted in the sentence context.

In all the sentences we ensured that there was no clause boundary directly following the embedded word, and that the onset of the following word matched the second syllable of the longer word. In this way, even allowing for co-articulation, the acoustic realization of the embedded syllable should be as similar as possible to the onset of the longer word. An example pair of sentences is shown below.

Short Word: *The soldier saluted the flag with his* ***cap*** *tucked under his arm.*

Long Word: *The soldier saluted the flag with his* ***captain*** *looking on.*

Acoustic Analysis and Alignment Points

The aim of these experiments was to compare subjects' interpretations of the paired sentences at different points in the speech stream. However we needed to ensure that these comparisons were made between stimuli containing a matched amount of acoustic information. Since we expected the duration of the target syllables to differ in our stimuli, it was necessary to create *alignment points* at phonemically equivalent positions in each sentence.

The first alignment point (*alp1*) was chosen to be at the offset of the first syllable (e.g. following /kæp/), a point at which the speech should be as ambiguous as possible. As expected, there were significant differences in the duration of the syllable prior to this alignment point, with syllables taken from short words being approximately 50ms longer than the equivalent syllable from a bisyllable[4]. Differences in responses to the paired stimuli at this point would therefore suggest that listeners are sensitive to acoustic differences between syllables from short and long words.

The second alignment point (*alp2*) was placed following the onset of the second syllable (/kæpt/). There were no differences in the duration of this section (*alp2-alp1*) in the two sets of stimuli. The third alignment point (*alp3*) was placed a fixed number of pitch periods into the vowel of the second syllable (/kæptuː/ or /kæptɪ/), with again no differences in the length of the section of speech between *alp3* and *alp2*. It is only at *alp3* that there is a phonemic difference between the two stimuli. Accounts of recognition which rely on mismatch between embedded words and longer competitors would therefore predict that identification of the short words be delayed until this point.

Experiment 1: Gating

In the gating task, speech is presented to subjects in fragments (gates) of progressively increasing duration.

[2] Other non-phonemic variables listed by Klatt that alter syllable duration include speech rate, discourse focus and phrase structure

[3] Since monosyllables generally occur more frequently in the language than bisyllables, pair-wise matching of frequency was not possible. However, across the set of stimulus pairs, these differences were not significant, t(39)=1.07, p>0.1

[4] syllable duration (monosyllables) = 291ms, syllable duration (bisyllables) = 242ms, t(39)=9.35, p<0.0001

Following each gate, subjects write down the word or words that they can hear. This allows us to investigate listeners' responses to increasing amounts of acoustic information. In the experiment, gates were set up at the three alignment points described earlier, with additional gates placed at 50ms intervals before *alp1* and after *alp3*.

Since we are looking for effects of differences in duration that may be contrastive only by reference to prior speech rate and phrase structure, subjects heard the complete onset of the sentence at each gate. Subjects were provided with an answer book containing the onset of each sentence up to the target word, their task being to identify the continuation of the sentence based on the speech they heard at each gate. The 40 pairs of stimulus sentences were randomly divided into two versions so that subjects heard only one of each stimulus pair. Twenty two subjects were tested on the two versions of the experiment.

Results and Discussion

The proportions of responses at different gates matching the short and long target words are shown in Figure 1. At early gates up to and including the offset of the first syllable (*alp1*), the majority of responses match the short target (e.g. CAP). Even at the first gate, 100ms before *alp1*, subjects hear enough speech to identify the first syllable.

However, there were significant differences in the proportion of short word responses depending on which of the pair of stimuli subjects were hearing. Across the three gates up to *alp1*, subjects made significantly more short word responses to short word stimuli than to long word stimuli (F1[1,20]=84.08, p<0.001; F2[1,36]=26.58, p<0.001]. The reverse pattern was also true, with significantly more long word responses being given to long word stimuli across the earliest three gates (F1[1,20]=6.55, p<0.05; F2[1,36]=4.50, p<0.05]. This difference in responses to the two sets of stimuli suggests that subjects are sensitive to the acoustic cues that differentiate between the syllables of short and long words.

Despite this difference, the recognition of embedded words still appears to be delayed relative to the identification of the longer words that they are embedded in. It is not until gate eight that subjects give as many correct responses to the short stimuli as to the long stimuli.

This delayed recognition appears to be caused by competition from longer word hypotheses, since at *alp2* (the onset of the second syllable) subjects gave many more long word responses to short word stimuli than at the previous gate. It is only when there is clear mismatch between the short word stimuli and the long target words (i.e. mismatch between *cap tucked* and *captain*) at *alp3* that subjects are able to reject the long word hypothesis. This suggests that the identification of monosyllabic words may be delayed until longer interpretations can be ruled out (cf. Grosjean, 1985), as simulated by TRACE.

However, it is unclear whether this result merely reflects a bias in the gating task. Since subjects must generate a response at each gate, it seems likely that they will try to produce a single word that encompasses all of the speech they can hear, instead of guessing the identity of speech they have yet to hear. So, at gates up to *alp1* where subjects are hearing speech from a single syllable, they will tend to produce monosyllabic words in response. This accounts for the large proportion of short word responses observed at these early gates. At *alp2* (onset of the second syllable)

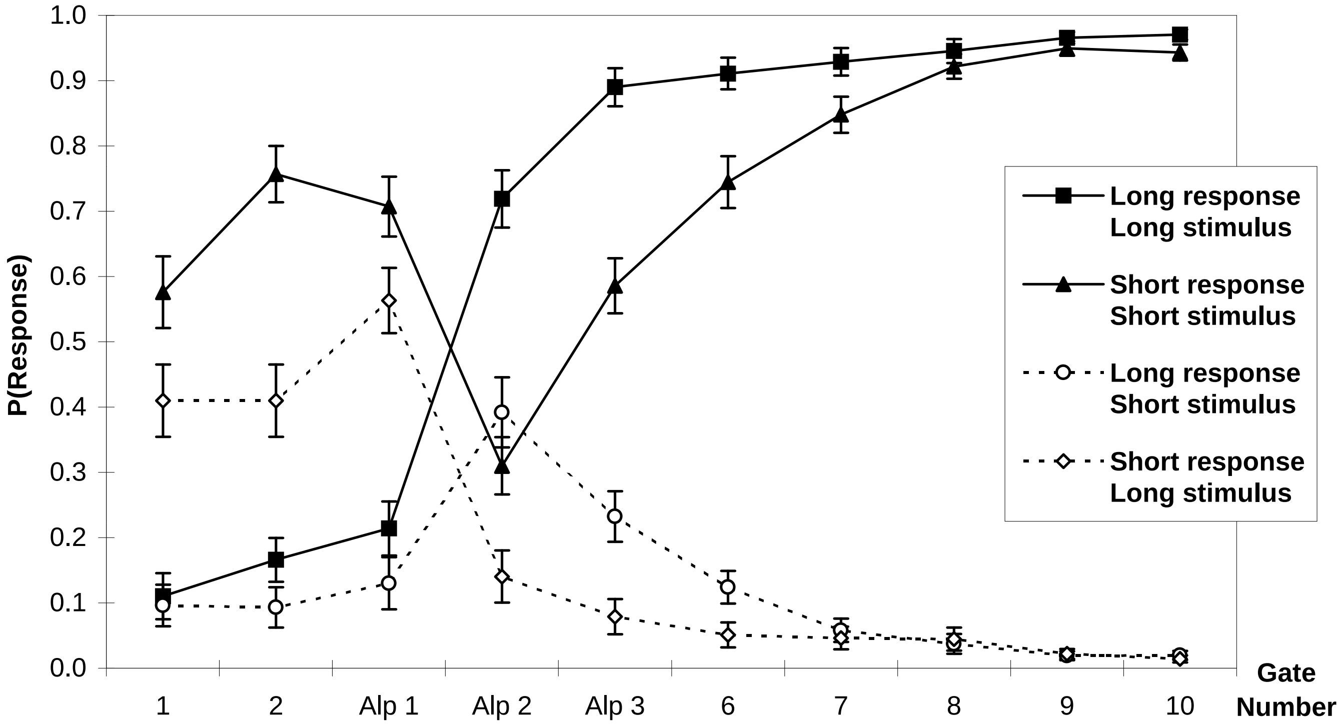

Figure 1: Experiment 1 - Gating. Proportion of responses matching the target words (CAP/CAPTAIN) for short and long word stimuli (*cap tucked/captain*). Error bars are 1 standard error.

instead of continuing to produce a monosyllabic response and guessing at the following word, subjects would tend to interpret the continuation as belonging to the same word, even where the following syllable is actually the onset of a new word. Such a bias towards giving a single word in response would increase the number of long word responses to short word stimuli. By reducing the number of short word responses that subjects give at *alp2* this single word bias could exaggerate the delayed recognition of embedded words.

Experiment 2: Cross-modal Repetition Priming

The goal of Experiment 2 was to investigate the recognition of the target words in our test stimuli using a task less susceptible to the biases suggested for gating. Instead we used a lexical decision task where subjects respond to a visual target following an auditory prime. Comparing reaction times following related and unrelated primes provides a measure of priming, the magnitude of which can be used to indicate the activation of different word hypotheses at a particular point in the speech stream (cf. Zwitserlood, 1989).

The same set of 40 paired test sentences from Experiment 1 were used as primes in this experiment, along with a third sentence in which the target words were replaced by an unrelated (although contextually viable), frequency-matched control prime. The prime sentences were cut off at probe positions equivalent to the gates used in the previous experiments, at which point one of the target words was visually presented, with subjects making a lexical decision response on a button box. The three prime conditions and two target conditions are shown in Table 1.

Table 1: Example primes and targets for Experiment 2 - Cross-modal priming. Primes and continuations following the sentence: *"The soldier saluted the flag with his..."*

Prime Type	**Prime Word** (continuation)	**Short Target**	**Long Target**
Short Test	***cap*** *(tucked under his arm)*	CAP	CAPTAIN
Long Test	***captain*** *(looking on)*	CAP	CAPTAIN
Control	***palm*** *(facing forwards)*	CAP	CAPTAIN

The prime and target conditions were rotated into a 6 version experiment, such that subjects were only presented with one condition for each item. In addition to the 40 test sentences, a set of phonologically-related non-word foils were included in each version (so that similarity between prime and target wasn't paired with a 'yes' response). Word and non-word fillers were also added to produce experimental versions in which 16% of all trials had a word target related to the auditory prime.

Experiment 2(a) - *alp1*

The initial experiment run using this method used sentence primes cut off at alp1 - the offset of the first syllable of the target words. Measures of priming were taken by comparing RTs to targets following test and control primes. The amount of facilitation for each of the prime and target conditions is shown in Figure 2.

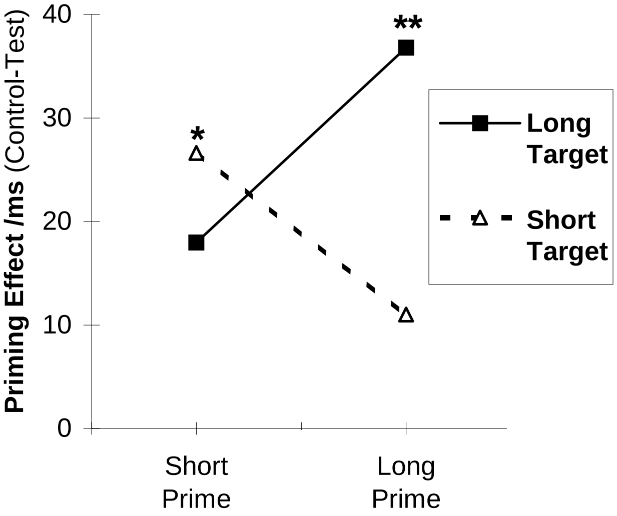

Figure 2: Experiment 2(a) - Magnitude of priming at *alp1* for short and long prime words (***cap*** *[tucked]*/***cap****[tain]*) and targets (CAP/CAPTAIN). * p<0.05; ** p<0.01 (control-test)

This pattern of results is very different to that observed in Experiment 1. Unlike in gating, there was no evidence of an overall bias towards short word interpretations at *alp1*. An ANOVA on the control-test differences showed no main effects of either prime or target type (i.e. no greater priming of CAP than of CAPTAIN) suggesting that the cross-modal priming task is less susceptible to the biases observed in gating. This may result from several differences between the two experiments; firstly in priming subjects are no longer required to explicitly identify the test stimuli, secondly in an on-line task subjects may be less inclined to interpret silence following the probe position as a word boundary.

Furthermore, the significant interaction between prime and target type (F1[1,57]=9.14, p<0.01; F2[1,31]=5.03, p<0.05) suggests that acoustic differences between stimuli containing short and long words, have a significant effect on subjects' interpretations before the stimuli diverge phonemically. This is confirmed by planned comparisons showing that priming is only significant where the word being heard at the probe position is identical to the visually presented target (i.e. the first syllable from *captain* primes CAPTAIN but not CAP and vice- versa).

From the results of this experiment, it seems that the claim made by McQueen et al. (1995), that lexical competition is necessary to account for the recognition of embedded words such as *cap* is premature. Priming results suggest that at the offset of an embedded monosyllable, listeners already have acoustic evidence to allow discrimination of syllables from short and long words.

Such results suggest a sensitivity to information in the speech stream that, although not phonemically contrastive, does differentiate between lexical items. We therefore use this as evidence that the processes of lexical segmentation and lexical access use sources of information, such as syllable duration, that may only be contrastive by reference to prior spoken context.

Experiment 2(b/c/d) - *alp2/alp3/gate7*

Despite the apparent utility of acoustic differences between the syllables of short and long words, on the basis of Experiment 2(a) we are unable to rule out the role of post-offset mismatch (and hence lexical competition) in the recognition of embedded words. In order to investigate how information that arrives after the offset of an embedded word affects recognition we decided to carry out further priming experiments using later probe positions. The positions chosen were the remaining alignment points from the gating experiment (*alp2* and *alp3*) as well as a probe point 100ms after *alp3* (equivalent to gate 7). Three separate priming experiments were carried out, one for each probe position, using the conditions shown in table 1.

As can be seen in Figure 3, the general pattern of priming effects in Experiment 2(a) were confirmed in all four experiments. An ANOVA on the control-test difference scores showed a highly significant interaction between prime and target type across all of the probe positions tested (F2[1,37]=28.95, p<0.001). However the exact pattern of priming effects observed in each experiment was not homogeneous, as suggested by a marginally significant 3 way interaction between prime, target and probe position (F2[3,113]=2.46, p=0.067) in the difference scores analysis.

One change in the priming effects is for short word targets at *alp2*, where subjects hear the onset of a syllable that continues to match the long target. At this probe position, priming of long word targets by long word primes is increased, whereas facilitation of short word primes by short word targets is reduced. This can be seen in the difference scores ANOVA at *alp2*, where in addition to the previously discussed interaction between prime and target type, there is also a significant main effect of prime type (F1[1,43]=5.73, p<0.05; F2[1,33]=5.77, p<0.05) and a marginally significant effect of target type (F1[1,43]=3.54, p=0.067; F2[1,33]=3.49, p=0.071). This difference in the priming effects is reminiscent of the results found in the gating experiment, where subjects produced many more long word responses to both types of stimuli at *alp2*.

Since we have already suggested, based on Experiment 2a, that cross-modal priming is less susceptible to the single word bias that was a potential confound in gating, this result suggests that post-offset information does indeed affect the activation of embedded words such as *cap*. More specifically, we have some evidence that the activation of the short word hypothesis in this experiment may be impaired by the continuing match between the onset of the following word and a longer competitor. This may also be responsible for the marginally greater priming for long targets observed in the ANOVA across all four experiments (F2[1,37]=3.71, p=0.062).

The effect of delayed mismatch between short word stimuli and long targets is also apparent in the results at later probe positions. It is only at *gate 7* (100 ms after *alp3*) that responses to both target words are significantly primed, not only by comparison with control primes, but also by comparison with related but mismatching primes. This pattern is similar to that observed in gating, where phonemic differences following *alp3* aids the rejection of alternative interpretations and allows the recognition of the short target words.

Discussion

The experiments we report here have shown significant

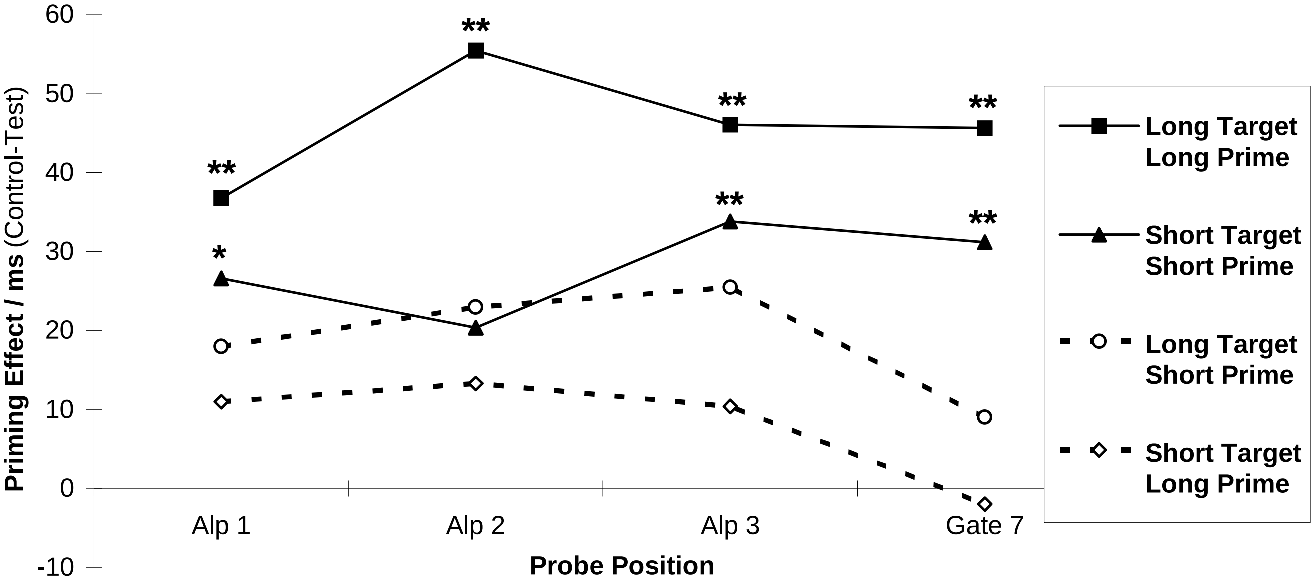

Figure 3: Experiment 2 - Magnitude of priming for short and long primes (***cap/captain***) with short and long targets (CAP/CAPTAIN) across 4 probe positions. * p[subject/item]<0.05; **p [subject/item]<0.01. (control-test).

differences in subjects' interpretations of syllables that come from short words or longer words in which these syllables are embedded. In particular, at the offset of an embedded syllable, reliable priming is only observed where the target matches the word from which the prime syllable has been taken.

This result suggests that left-embedded words are not as ambiguous as has previously been suggested in the literature. Acoustic cues to word length are present in the speech stream, and are used by listeners during the recognition of connected speech. Although the exact nature of these cues is still under investigation, it seems likely that differences in syllable duration play an important role in the recognition of embedded words.

However, given the wide variation in syllable duration between different utterances and speakers, it is likely that such cues can only be contrastive through comparison with preceding spoken context. Models of lexical segmentation and speech perception need to be constructed to investigate the mechanisms by which listeners adapt to variations in speech rate encountered in normal conversation (see Abu-Bakar and Chater, 1994, for some related work modeling rate adaptation in phoneme categorization).

Despite the apparent utility of these acoustic cues to word length, we have also seen that following context does affect the recognition of embedded words, as would be predicted by TRACE and other models. The experimental stimuli used here where continuations match a longer competitor may present particular problems for the recognition system. Future work is investigating the role of continuations that are either phonotactically or lexically non-viable in order to determine whether the role of following context necessarily entails lexical level competition or can be accounted for within a system mapping directly from the speech stream to a distributed representation of form and meaning (cf. Gaskell and Marslen-Wilson, in press).

Acknowledgments

This research was supported by an EPSRC post-graduate research studentship number 94700590. We wish to thank members of the Birkbeck Centre for Speech and Language for useful discussions.

References

Abu-Bakar, M., & Chater, N. (1994). Distribution and frequency: Modeling the effects of speaking rate on category boundaries using a recurrent neural network. In Ram & Eiselt (Eds.), *Proceedings of the sixteenth annual conference of the cognitive science society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baayen, R. H., Pipenbrook, R., & Guilikers, L. (1995). *The Celex Lexical Database* (CD-ROM). Linguistic Data Consortium, Univ. of Pennsylvania, Philadelphia, PA.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics*, 44, 395-408.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1994). Lexical segmentation: the role of sequential statistics in supervised and unsupervised models. In Ram & Eiselt (Eds.), *Proceedings of the sixteenth annual conference of the cognitive science society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Content, A., & Sternon, P. (1994). Modeling retroactive context effects in spoken word recognition with a simple recurrent network. In Ram & Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. Journal of Experimental Psychology: Human Perception and Performance, 14(1), 113-121.

Gaskell, M. G., & Marslen-Wilson, W.D. (in press). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes*.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, 38(4), 299-310.

Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.

Klatt, D. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59(5), 1208-1221.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, 51(6), 2018-2024.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics*, 39, 155-158.

Marslen-Wilson, W. D. (1996). Abstractness and combination: the morphemic lexicon. Manuscript.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes*, 10, 309-331.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, 62(3), 715-719.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing words without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, 63(1), 234-245.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32, 25-64.