LEXICAL SEGMENTATION AND AMBIGUITY: INVESTIGATING THE RECOGNITION OF ONSET-EMBEDDED WORDS

Matt. H. Davis, William D. Marslen-Wilson
MRC Cognition and Brain Sciences Unit, Chaucer Road, Cambridge, UK
and **M. Gareth Gaskell**
Department of Psychology, University of York, York, UK.

ABSTRACT

The lack of acoustic markers of word boundaries in connected speech may create temporary ambiguities between words like *cap* and the start of longer words like *captain*. These ambiguities have motivated models of spoken word recognition in which lexical competition allows information after the end of an embedded word to assist identification. We review the results of priming experiments demonstrating that additional acoustic cues assist listeners in distinguishing embedded words from longer competitors. We present a recurrent network model in which acoustic information and following context combine in the identification of onset-embedded words. Simulations show an activation profile consistent with the cross-modal priming data suggesting that recurrent networks can be used to model the identification of onset-embedded words.

1. INTRODUCTION

Connected speech contains few reliable acoustic markers of word boundaries [1]. Although non-lexical cues or strategies may allow segmentation prior to lexical access [2,3] models of spoken word recognition typically incorporate mechanisms so that lexical identification can contribute to word segmentation. For instance, the cohort model [4] proposes that since words can be identified at their uniqueness point (which is often before their acoustic offset) recognition can assist the detection of word boundaries.

However, sequential recognition accounts of segmentation are challenged by words that do not become unique before their offset. For a word like *cap*, (embedded in *captain, captive, capsule*, etc.) it may not be possible to rule out longer competitors before the end of the word [5]. Sequences that contain ambiguities between onset-embedded words and longer competitors may therefore prove problematic for sequential recognition accounts of word segmentation.

Consequently, many models of spoken word recognition [6,7] incorporate processes of competition between non-aligned lexical hypotheses. Inhibitory connections between units spanning word boundaries allow the delayed identification of embedded words. For an example sequence *cap fits*, hearing /f/ after /kæp/ rules out all longer lexical items. Networks that incorporate lexical competition can use following context to increase the activation of onset-embedded words. The delayed recognition predicted by these accounts have been reported in gating experiments [8] while effects of following context have also been observed in word-spotting [3,9].

2. EMBEDDED WORDS AND AMBIGUITY

The need for delayed recognition of onset-embedded words is based on an assumption that there is ambiguity between *cap* and the first syllable of longer words like *captain*. However, evidence from acoustic phonetics suggests systematic differences between the syllables of monosyllabic and bisyllabic words. For example, Lehiste [1] reports significant shortening of the syllable [slɪp] in words like *sleepy* and *sleepiness*. Differences are also observed in segments adjacent to word boundaries [10], through co-articulation from following words and from prosodic boundaries.

In a series of experiments Davis, Marslen-Wilson and Gaskell [11] investigated the degree of ambiguity between embedded words and longer competitors in sentences. Stimulus sentences for these experiments contained monosyllabic words or frequency-matched bisyllables that contained the monosyllable as the initial syllable (such as the pair *cap* and *captain*). Forty pairs of words were placed in non-biasing sentence contexts. Continuations of the short word stimuli formed a lexical garden-path with the onset of the following word matching the onset of the second syllable of the longer word (e.g. *cap tucked*). Thus co-articulatory influences from the following word should not provide information to distinguish between embedded words and longer competitors. Example sentences are shown in Table 1.

Markers were placed in these sentences at phonemically aligned positions: at the offset of the initial syllable (1. after /kæp/), after the onset of the following syllable (2. /kæpt/), and in the vowel of the second syllable (3. /kæptʌ/ or /kæptɪ/). Acoustic analyses showed that the duration of the embedded syllable /kæp/ was significantly longer in the short words. A gating study showed that responses at gates up to and including probe 1 differed for the two sets of sentences [11].

Prime Type	**Prime Word** continuation	**Short Target**	**Long Target**
Short Test	***cap***[1] *t*[2]*u*[3]*ck*[4]*ed under his arm*	CAP	CAPTAIN
Long Test	***cap***[1]***t***[2]***ai***[3]***n***[4] *looking on*	CAP	CAPTAIN
Control	***rifle*** *by his side*	CAP	CAPTAIN

Table 1: Primes, targets and probe positions (1-4) following: *"The soldier saluted the flag with his..."*

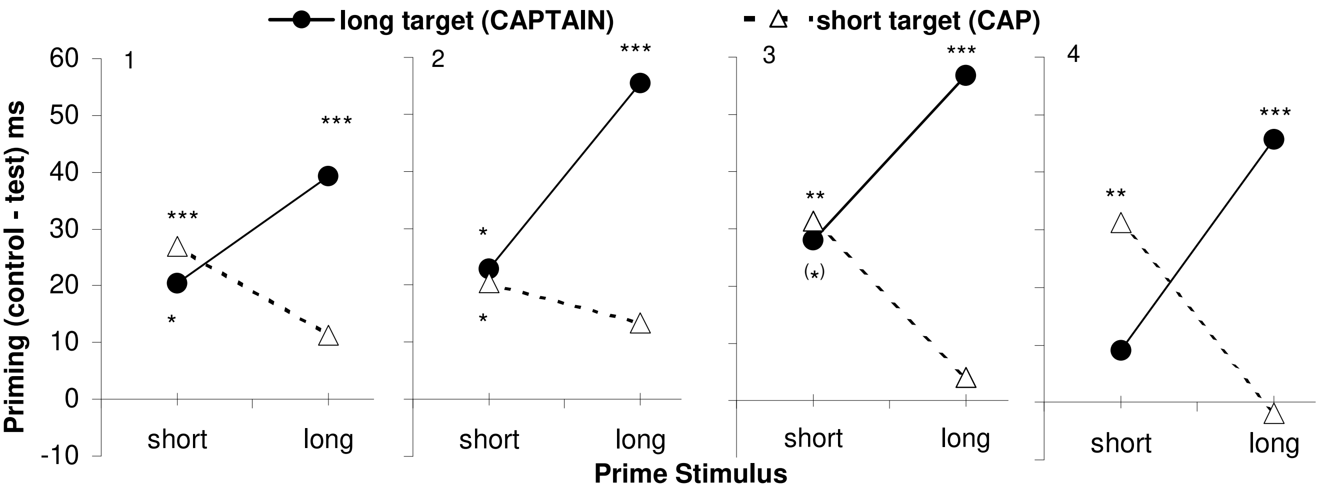

Figure 1: Magnitude and significance of priming from short (*cap tucked*) and long *(captain)* primes to short (CAP) and long (CAPTAIN) targets in at probe position 1 (/kæp/), 2 (/kæpt/), 3 (/kæptʌ/ or /kæptɪ/), 4 (100ms after probe 3)

Four cross-modal, repetition-priming experiments were carried out to investigate the activation of onset-embedded words and longer competitors. In each experiment, the magnitude of priming for short and long target words was used as a measure of lexical activation; comparing responses following test prime sentences to a control sentence containing an unrelated (but contextually viable) word. Each experiment used sentence primes cut off at the probe position to track lexical activation across critical portions of the speech stream. The four experiments tested the three alignment points described previously, as well as an additional probe position 100ms after position 3. Prime and target conditions are shown in Table 1.

Priming effects by prime and target type (with significance levels marked) are shown in Figure 1. At all probe positions ANOVAs on control-test differences showed a significant interaction between prime and target type ($p<.05$). At the first and last probe position there were no main effects of prime or target type ($p>.1$). This indicates that greater priming was observed where the prime and target contained the same word compared to the case in which they were only phonologically related. This is unsurprising at later probe positions – participants heard enough of the test prime sentences to distinguish embedded words from longer competitors and vice-versa. However the cross-over interaction at the first probe position shows that this effect was also apparent at the end of the embedded syllable (/kæp/). This result provides a clear indication that some acoustic cue or cues are available in the initial syllable of the critical words that bias the activation of embedded words and longer competitors.

Despite the effectiveness of this acoustic cue, information after word offset does affect the perception of embedded words. Longer competitors remain active at later probe position as shown by priming of long word targets at probe positions 2 and 3. The activation of longer competitors is also indicated by marginally greater overall priming for long targets at these positions ($p<.1$). Lexical garden-path continuations support the continued activation of longer lexical items in short word stimuli. This activation is suppressed at the final probe position where there is sufficient mismatch in the input to rule out longer lexical items.

3. RECURRENT NETWORK SIMULATIONS

Sequential recognition accounts of lexical segmentation have been criticised for their inability to identify onset-embedded words. Neural network simulations illustrate this limitation. Simple recurrent networks trained to map from a sequence of speech segments to a representation of the current word in the speech stream [12,13] are unable to distinguish embedded words from longer competitors since they cannot backtrack and revise their interpretation of previously ambiguous input. Although acoustic cues that distinguish between embedded words and longer competitors might reduce the severity of this ambiguity, evidence of post-offset activation of longer competitors is also problematic for accounts of segmentation in which words must be identified before their acoustic offset.

It is argued that this limitation of sequential recognition accounts necessitates the inclusion of inhibitory competition between lexical units in models of spoken word recognition. However recurrent network simulations in which the output representation incorporates information about previous words in the input [14,15] have are capable of using following context to revise interpretations of prior words. These simulations demonstrate that the computational properties required for the identification of onset-embedded words can be achieved in a connectionist learning environment, without the inclusion of direct inhibitory connections between lexical units. However, these simulations still fall short of providing an account of the experimental data presented here; it is necessary to show that, when provided with appropriately structured input representations, recurrent networks can simulate the combination of acoustic and lexical influences on segmentation suggested by the priming data.

The model described here is based on the simulations reported by Davis, Gaskell and Marslen-Wilson [15]. It consists of a simple recurrent network trained to activate a localist representation of all the lexical items contained in sequences of 2 to 4 words. Since the network must maintain a representation of all the words in an utterance, it is able to use following context to identify embedded words. The training set for these simulations was an artificial language of CVC syllables represented over a reduced set of phonetic features. The network's vocabulary included bisyllabic words with onset-embeddings (e.g. *cap* and *captain*).

An important difference between these simulations and those reported previously [15] is the inclusion of three input units representing additional, non-phonemic cues to word length. Although intended to be analogous to the duration differences observed between syllables in short and long words, the input representations used here are not specifically durational, and could be considered analogous to any non-phonemic acoustic cue which can distinguish between syllables in short and long words in a noisy and contextually dependant fashion. The three additional inputs were activated during each vowel to indicate the length of the current syllable; monosyllables being associated with longer durations. Since syllable duration also varies with the overall rate of an utterance, two out of the three duration codes were chosen based on a randomly assigned speech rate for each sequence in the training set. Thus utterances spoken at a fast rate would use the shortest code for bisyllables and the middle code for monosyllables while slower utterances would use the middle code for bisyllables and the longest code for monosyllables. Consequently, the network must use prior context to disambiguate syllables at the middle duration. To reduce the reliability of this additional input, the ambiguous middle code replaced the fast and slow duration codes in 20% of words.

The network architecture used here is shown in Figure 2. Additional units not shown in this diagram were trained to predict the input at the next time step and to output a copy of the current and previous input [2]. Ten networks were trained using back-prop from random initial weights on 500,000 sequences.

To compare model and data we assumed that the magnitude of priming can be directly related to activation at the lexical output. Ten fully trained networks were tested on lexical garden path sequences and long words containing onset-embedded words. In all test sequences the critical words were presented as the second word in a sequence and with the middle duration code such that the network has to use prior context to disambiguate the input. Since network output can only be tested at the offset of each segment, probe positions correspond to the phonemes present at each marked position in the test stimuli.

At all the probe positions tested in the network, ANOVA showed a significant interaction between lexical unit and input sequence (all $p<.001$). This interaction shows that the network succeeds in

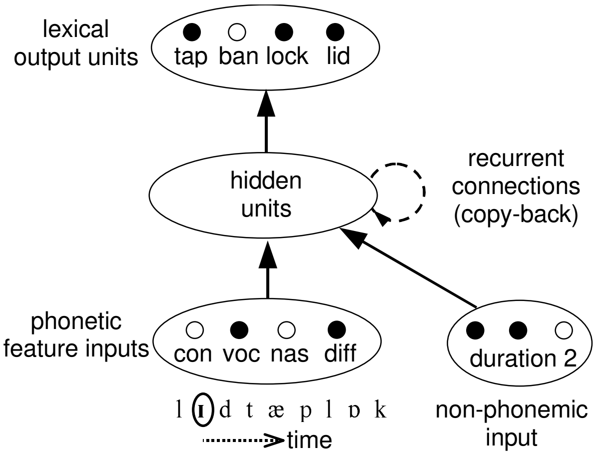

Figure 2: SRN trained to activate a representation of all the words in a sequence using phonetic and non-phonemic input

simulating one of the main results from the priming experiments. At the offset of the embedded syllable, greater lexical activation is observed for appropriate interpretations of the input sequences. This finding could not be simulated by a model in which lexical activation depends solely on phonemic information. In such a model, this interaction would only be observed at probe positions where the input sequences differ phonemically (position 3 or 4).

As was observed in the experimental data there was no effect of stimulus type ($F<1$) at the first probe position, though the network did show a marginally significant effect of lexical unit ($p<.1$). Unlike the experimental data, the network predicts marginally greater activation for short words at the earliest probe position. However, since lexical decision responses were faster for the short word stimuli, there may be other reasons why greater priming of short words was not observed in the priming experiments.

At the second and third probe position, the network predicts greater activation for long lexical items (probe 2, $p<.001$; probe 3, $p<.001$) equivalent to the effect of target type observed in the experimental data. The presence of lexical garden-paths in the short word stimuli leads to continued activation of long lexical hypotheses. At the final probe position, although the strength of this main effect is reduced it remains significant ($p<.01$). Thus compared to the experimental data, the network predicts longer lasting disruption from lexical garden-paths though these networks are capable of identifying the onset-embedded words.

4. DISCUSSION

We have described and simulated experimental results that illustrate two processes involved in the identification of onset-embedded words. Firstly, priming data shows that some acoustic cue or cues allow the perceptual system to distinguish embedded words from longer competitors before their acoustic offset. The interaction between prime and target length observed in the priming data are correctly simulated in a model with additional inputs that provide a non-

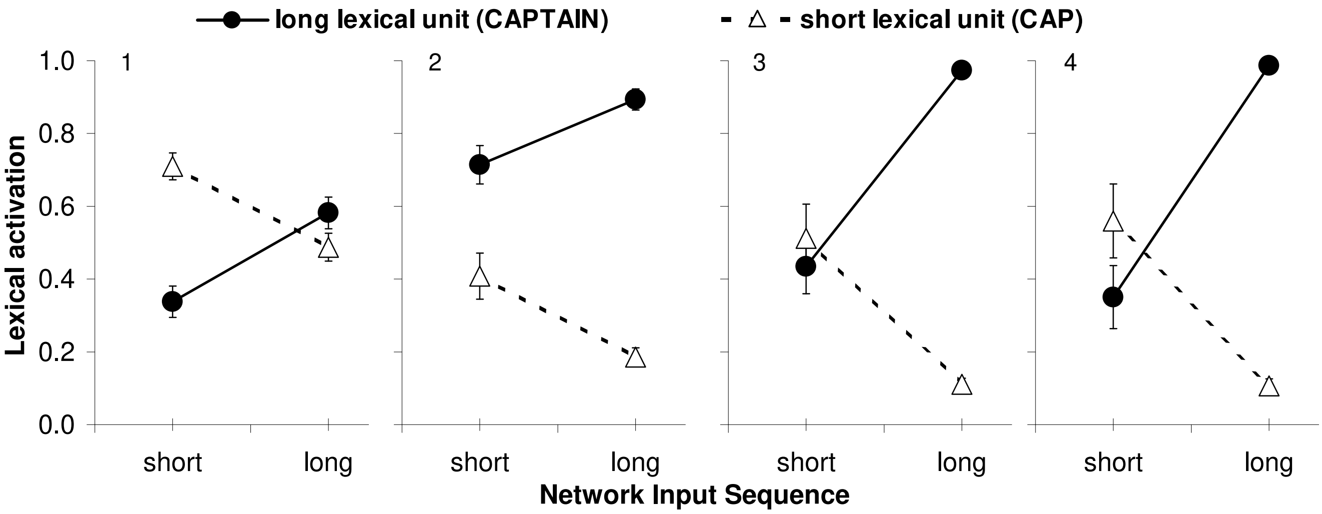

Figure 3: Network activation of short (CAP) and long (CAPTAIN) lexical units for sequences containing short (*cap tap*) and long *(captain*) words. Probe position 1 (/kæp/), 2 (/kæpt/), 3 (/kæptæ/) or (/kaptɪ/), 4 (/kæptæp/) or (/kaptɪn/).

phonemic cue to the length of a word. Secondly, we have shown that in spite of these acoustic cues, for lexical garden-path sequences (e.g. *cap tucked*), words like *captain* remain active after the offset of the embedded word. A recurrent network trained to preserve the activation of lexical units across sequences of words shows the same property with long words remaining active in response to lexical garden-path sequences. These results indicate that recurrent neural networks can be used to model the integration of lexical and acoustic cues in the recognition of embedded words without requiring direct inhibitory connections between lexical items. Further simulations using recurrent networks are merited to investigate whether they can simulate other experimental data on the segmentation and identification of words in connected speech.

5. ACKNOWLEDGEMENTS

This work was supported by EPSRC studentship 94700590 and by an MRC program grant to W. Marslen-Wilson and L. K. Tyler. Correspondence to: Matt Davis, MRC-CBU, 15 Chaucer Road, Cambridge, CB2 2EF, UK. Email: matt.davis@mrc-cbu.cam.ac.uk.

6. REFERENCES

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

[2] Cairns, P., Shillcock, R., Chater, N., Levy, J. (1997). Bootstrapping word boundaries: a bottom-up corpus based approach to speech segmentation. *Cognitive Psychology*, **33**, 111-153.

[3] 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.

[4] Marslen-Wilson, W.D. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

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

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

[7] Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, **52**, 189-234.

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

[9] McQueen, J.M., Norris, D. & Cutler, A. (1994). Competition in spoken word recognition: spotting words in other words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **20**(3), 621-638.

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

[11] Davis, M.H., Marslen-Wilson, W.D. & Gaskell. M.G. (1997) Ambiguity and competition in lexical segmentation. In Shafto, M. & Langley, P. (Eds.) *Proceedings of the 19th Conference of the Cognitive Science Society*. LEA: Hillsdale, NJ.

[12] Norris, D., A dynamic-net model of human speech recognition in Altmann, G.T.M. (Ed.) *Cognitive Models of Speech Processing*, MIT Press: Cambridge, MA.

[13] Gaskell, M.G. & Marslen-Wilson, W.D. (1997). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes*, **12**, 613-656.

[14] Content, A. & Sternon, P. (1994) Modelling retroactive context effects in spoken word recognition with a simple recurrent network. In Ram, A. & Eiselt, K. (Eds.) *Proceedings of the 16th Conference of the Cognitive Science Society*, LEA: Hillsdale, NJ.

[15] Davis, M.H., Gaskell, M.G. & Marslen-Wilson, W.D. (1997). Recognising embedded words in connected speech: Context and competition. In Bullinaria, J., Glasspool, D., Houghton, G. (Eds) *Proceedings of the 4th Neural Computation and Psychology Workshop*, Springer-Verlag: London.