Recognising Embedded Words in Connected Speech: Context and Competition

Matt H. Davis, M. Gareth Gaskell and William Marslen-Wilson
Centre for Speech and Language
Birkbeck College, London

Abstract

Onset-embedded words (e.g. *cap* in *captain*) present a problem for accounts of spoken word recognition since information coming after the offset of the embedded word may be required for identification. We demonstrate that training a simple recurrent network to activate a representation of all the words in a sequence allows the network to learn to recognise onset-embedded words without requiring a training set that is already lexically segmented. We discuss the relationship between our model and other accounts of lexical segmentation and word recognition, and compare the model's performance to psycholinguistic data on the recognition of onset-embedded words.

1   Introduction

A significant problem for both human and machine speech recognition is how to determine the location of boundaries between lexical items or words. Connected speech does not contain gaps between words analogous to the spaces in written text [1], nor are there any explicit cues that reliably mark the position of boundaries between words [2]. Accounts of how listeners segment the speech stream during spoken word recognition can be divided into two broad classes. The first is based on the use of sub-lexical cues that can identify word boundaries prior to, or in the absence of, lexical access to words in the speech stream. Examples of cues that have been suggested in the literature include, prosodic [3], metrical [4] or phonotactic [5] regularities, or some combination of these [6,7]. However, since these cues are unable (either singly or in combination) to reliably identify all word boundaries, a second class of account has also been proposed suggesting that for adult listeners, the recognition of individual words contributes to lexical segmentation. We will focus on this second class of account here.

In theories of lexical access, most notably the cohort model [8], the recognition of connected speech is argued to proceed in a maximally efficient manner, i.e. words are identified as soon as they become uniquely specified in the speech stream. By this account, many words can be recognised before their offset, and word boundaries can be predicted for the offset of the current word. However, the presence of large

numbers of words which are not unique at their offset [9] presents a challenge to lexical accounts of segmentation. At the offset of a word such as *cap*, the recognition system may be unable to distinguish this word from longer competitors (such as *captain* or *captive*) and may therefore be unable to reliably locate a word boundary. This seems to require that the recognition of onset-embedded words be delayed until after a word boundary, when longer competitors can be ruled out. This is supported by experiments showing that monosyllables are commonly not recognised until after their acoustic offset [10,11]. Such results have been taken as evidence that no sequential recognition system is capable of recognising onset-embedded words.

1.1 Connectionist accounts of spoken word recognition

Connectionist models of spoken word recognition illustrate many of the problems associated with different accounts of lexical segmentation. Simple recurrent networks [12], trained to map sequences of phonetically coded segments to a lexical representation of the current word [13, 14], show maximally efficient recognition of words in connected speech. During recognition, the network produces parallel activation of lexical representations for all the words that match the current input. So for the input sequence /kæptɪ/ (matching both *captain* and *captive*), a network trained to represent the lexical identity of words in the input will activate a blend of the representations of both words. In producing this parallel activation, the network activates the arithmetic mean of the output representations of all the words that match the speech input. Consequently, the network only fully activates a single lexical representation when the speech stream uniquely identifies a word.

However, since these networks are trained to activate a representation of the current word in a sequence, they are unable to correctly identify onset-embedded words such as *cap*. For these items, the speech stream only uniquely identifies such words after a lexical boundary, by which point the network will be activating a representation of the following word. Therefore at no point in the speech stream will the network fully activate the lexical representation of an onset-embedded word.

An alternative account of the time course of lexical activation in spoken word recognition is used in interactive activation models such as TRACE [15]. In these models, mutually inhibitory connections between lexical units that share segments produce partial activation of multiple candidates that match the speech input. Since these inhibitory connections cross potential word boundaries, the network can use following contexts (that rule out longer competitors) to account for the delayed recognition of onset-embedded words.

However, the architecture used in TRACE is computationally inefficient since it relies on duplicating units and connections to code for different time slices. Consequently Shortlist [16] combines the bottom-up activation of lexical candidates in a recurrent network with a competition network 're-wired on the fly' to account for the recognition of onset embedded words. Effective as this hybrid model may be [17], it remains of interest whether the approach used in the recurrent network simulations (where effects of competition are a result of the network evaluating the conditional

probability of different words at each point in the speech stream) can also be used to account for the delayed recognition of onset embedded words.

Content and Sternon [18] show that adding a set of output units representing the previous word in a sequence (as well as the current word) allows a simple recurrent network to recognise onset-embedded words. The additional output task ensures that the network continues to represent hypotheses regarding the identity of previous words in the speech stream, and is therefore able to revise its output using following context. This allows the network to display appropriately delayed recognition of onset embedded words. However as in previous SRN simulations, the target vector is changed at each word boundary. Thus the training regime requires specification of the location of word boundaries in the speech stream, as well as the identity and order of occurrence of the lexical items divided by these boundaries. This is unrealistic since it assumes that the language learner has access to a lexically segmented corpus in which correspondences between the speech stream and lexical items are already known. The goal of the work that we shall be reporting here is to investigate whether simple recurrent networks can be trained to lexically segment speech and recognise onset-embedded words *without* requiring a pre-segmented training set. We assume that in the early stages of vocabulary acquisition children are mapping whole utterances onto their interpretation of the world.

2   Learning to segment connected speech

The simple recurrent networks [12] investigated here were given the task of mapping sequences of phonemes to a representation of all the words contained in those sequences [19]. To allow easier interpretation of the network's output, this representation is composed of localist lexical units, each representing a word in the network's vocabulary. However, in contrast to previous simulations, the training set does not contain any explicit information about which segments in the speech stream map onto lexical items (i.e. the locations of word boundaries is not given). Nor is there any information about the order in which words occur in the training sequences. The target activation for the network remains static throughout an sequence; the network being trained to activate the lexical units for all the words in the input. A psychological account of this mapping is that the network is building an interpretation of each sequence of lexical items. By this account, the network must preserve a representation of words that have already been identified, as well as activating lexical representations for words as they occur in the input.

Since the network is unable to activate units representing the final words in a sequence until those words have been presented in the input, we cannot expect the network to learn the training set to perfection. However, as in networks learning to predict the next segment in an utterance, distinctions between the task on which the network is trained, and the performance of the network when tested may help provide a fully elaborated psychological account [6]. In the case of the networks investigated here, the immediate task for the network is to associate strings of phonemically coded segments with a representation of the lexical items contained in that sequence. During testing however, we will be interested in comparing the time-course of

activation of individual lexical items with available psycholinguistic data on the recognition of words in connected speech. The architecture of the network and the training regime are illustrated in Figure 1.
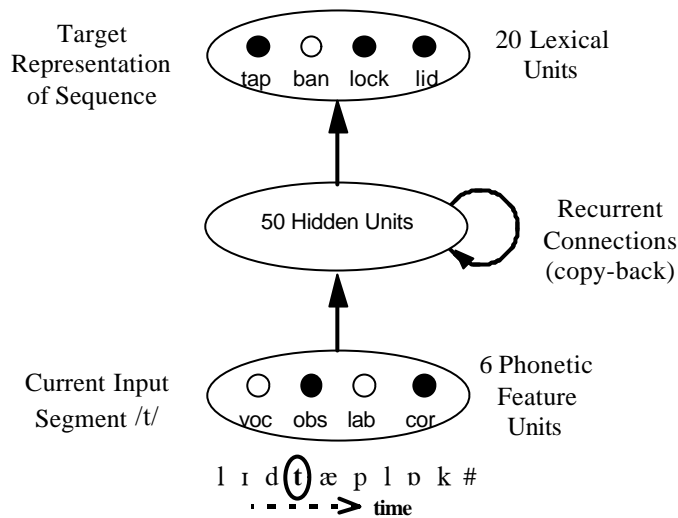

Figure 1: Simple recurrent network architecture used for these simulations showing a snapshot of training activations during the segment /t/ in the sequence "*lid tap lock*".
Throughout each training sequence, the target for the network is to activate a representation of all the words in that sequence, not just the current word.

An important issue in justifying any psychological model involving supervised learning is to elaborate where the training information comes from. As in other models of word learning [20] we assume that vocabulary acquisition involves learning a mapping from form to meaning. However, in contrast to these accounts (and previous models of spoken word recognition [13,14]) we do not assume that correspondences between the speech stream and lexical or semantic representations are available to the learner on a one to one basis. The assumption made in this model is that a substantial part of the problem of vocabulary acquisition is to extract these correspondences from experience of spoken words and the meanings of sequences of those words [21]. The specific implementation of this assumption - that all words in a sequence have lexical representations as targets during training - is less theoretically vital. It is likely that only a subset of the words in any utterance have an obvious interpretation during acquisition. Nonetheless the idea that correspondences between the speech stream and lexical/semantic representations are learnt from experience of sequences of words where meanings must be assigned to lexical items (instead of being pre-specified) is an important part of this model.

2.1 Training sets

For all the simulations reported here, the training set was constructed from an artificial language containing 7 consonants and 3 vowels coded over a set of 6 phonetic features [22]. These segments were placed in a CVC syllable template used to create a vocabulary of 20 lexical items. Of these 20 words, 14 were monosyllables and 6 bisyllables. To allow investigation of the time course of recognition, lexical items varied in the point at which they became unique from all other words in the networks vocabulary. The simulations included 'cohort' pairs such as *lick* and *lid*, that share the same onset and become unique on their final segment, as well as two pairs of onset-embedded words (e.g. *cap* and *captain*) where the monosyllable is not uniquely identifiable until following context rules out longer competitors. There were also two pairs of offset-embedded words (*lock* and *padlock*) to allow comparison of the network's sensitivity to preceding and following context in the recognition of embedded words. These vocabulary items are shown in Table 1.

Type	Word	Phonology	Word	Phonology
Onset-Embedded Bi.	captain	/kæptɪn/	bandit	/bændɪt/
Offset-Embedded Bi.	topknot	/topnot/	padlock	/pædlok/
Non-Embedded Bi.	landed	/lændɪd/	picnic	/pɪknɪk/
Onset-Embedded Mono.	cap	/kæp/	ban	/bæn/
Offset-Embedded Mono.	knot	/not/	lock	/lok/
Cohort Monosyllable	dot	/dot/	dock	/dok/
	lick	/lɪk/	lid	/lɪd/
Non-Embedded Mono.	tap	/tæp/	bat	/bæt/
	knit	/nɪt/	cat	/kæt/
	pot	/pot/	bid	/bɪd/

Table 1: Vocabulary items used in these simulations

These 20 words were randomly selected without replacement to create sequences of between 2 and 4 words in length, separated by a boundary marker (zero input and output vectors). No attempt was made to capture higher-order distributional regularities such as syntactic or constituent structure, however a subset of word pairs were held back during training to allow testing of the networks generalisation performance. These sequences of were presented to a simple recurrent network (6 inputs, 50 hidden units with copy-back connections, 20 outputs) trained to activate lexical units for all the words in the current utterance. Weights were updated by the standard back-propagation algorithm following the presentation of every input segment (learning rate = 0.02, no momentum, cross-entropy error).

In preliminary simulations we found that changes to the bias weights on the output units were considerably larger than those to weights connecting the output and hidden units. This is caused by repeated weight updates with the same target pattern. To allow larger learning rates in these simulations, bias weights were

disconnected from the output units in subsequent simulations[*]. Training time apart, results were comparable to simulations that included bias weights to the output units.

2.2 Results

Ten simulations were carried out using the network architecture and training regime described above. Each network was trained on 500 000 sequences using small, random initial weights and different seeds for generating the training sequences. Figure 2 shows the activation of target words for an example sequence averaged over 10 fully trained networks. As can be seen in the graph, the network partially activates words as segments are presented at the input. Words become fully activated when they are uniquely specified in the speech stream, and once identified remain active until the end of the sequence. Such behaviour indicates that the network has learnt to lexically segment the speech stream by associating sections of the speech with the corresponding lexical units. During training, the network was not provided with explicit cues to the location of word boundaries or with information about which segments make up individual lexical items. However, by generalising its experience of different sequences, the network has learnt correspondences between the speech stream and lexical items. This is illustrated by the network's identical pattern of performance on sequences that were held back during training.
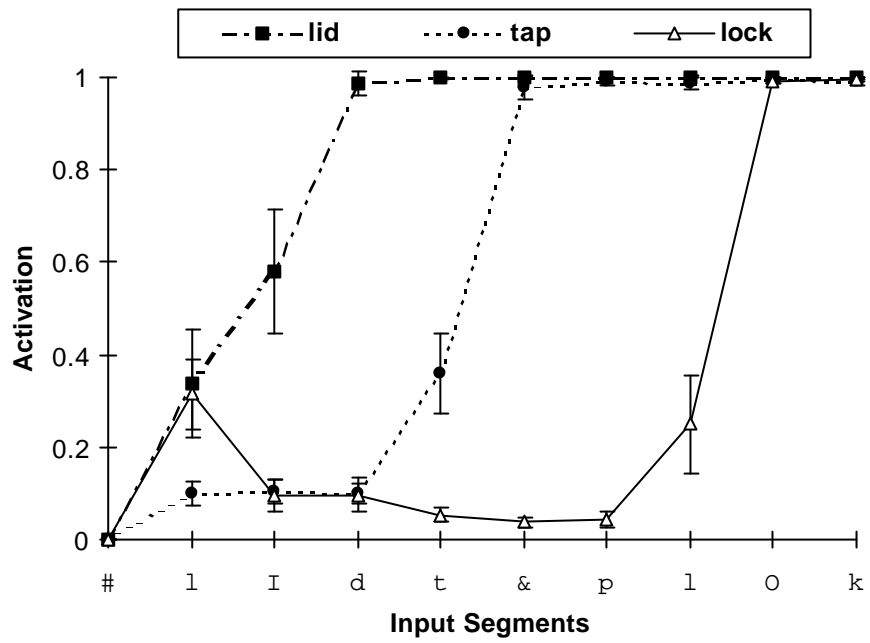
[*] Thanks to Gary Cottrell for suggesting this

Figure 2: Activation of target words during the sequence *"lid tap lock"*.
The network activates words as they are presented in the input and preserves their
activation until the end of the sequence. Error bars are 1 standard deviation.

2.2.1. Onset-embedded words

A systematic comparison of the partial activation observed for lexical items with different competitor environments is shown in Figure 3. The left hand chart shows the pattern of activation observed for items with cohort competitors (in this case *lick* and *lid*). As can be seen, output activations in the network approximate the conditional probabilities of partially matching lexical candidates [23]. Thus at the onset of *lid* (where three candidates match the input) each competing word is activated to just over 0.3. On presentation of the second segment, when two candidates match, each item is activated to approximately 0.5. It is only at the offset of *lid* that full activation is obtained at the appropriate lexical unit. Recent work using cross-modal semantic priming [24] is consistent with this account since the magnitude of priming observed for ambiguous word fragments is proportional to the conditional probability of the prime word in that cohort environment.
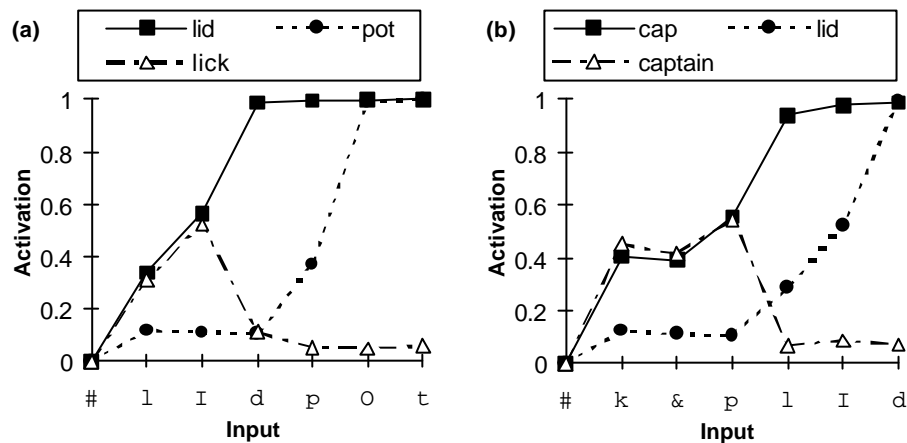

Figure 3: Activation of targets and competitors
(a) Cohort competitors (*lid/lick*) during the sequence *"lid pot"*
(b) Embedded words (*cap/captain*) during the sequence *"cap lid"*

The pattern of activation for cohort members is repeated almost identically in Figure 3(b) for onset-embedded words. At the offset of the monosyllable, the matching lexical items (*cap* and *captain*) are equally activated. It is only at the onset of the following word (*lid*) that the network receives disambiguating input (since the input mismatches with *captain*) and fully activates the target word *cap*. The network displays the optimal pattern of activation for ambiguous inputs, regardless of whether the ambiguity is resolved within a lexical item or following a word boundary. Such behaviour is comparable to experimental results from gating [10] since the network predicts that the identification of onset-embedded words requires post-offset information. Efficient recognition of embedded words like *cap* therefore does not require mutually inhibitory connections between competing lexical items.

These simulations support an account of segmentation in which post-offset mismatch with longer competitors plays an important role in the recognition of embedded words. But what is the pattern of activation displayed by the network when such mismatch is absent? Figure 4 shows the network's response in two such cases. The first example (4a) is where the network is presented with an item that contains a word embedded at the onset (for example *captain*). In this case the network strongly activates the longer competitor following the offset of the embedded word. However, in the 'lexical garden path' case in Figure 4(b), where the onset of the continuation matches the longer competitor, (e.g. *cap tap*) the network is still able to revise its lexical hypotheses in response to mismatch between the speech stream and the longer competitor. Further simulations showed that the network retained the ability to correctly resolve these lexical garden paths even when such sequences were not presented to the network during training. The disruption produced at the onset of the second word (*tap*) has also been confirmed in gating and cross-modal priming experiments using similar stimuli [25].
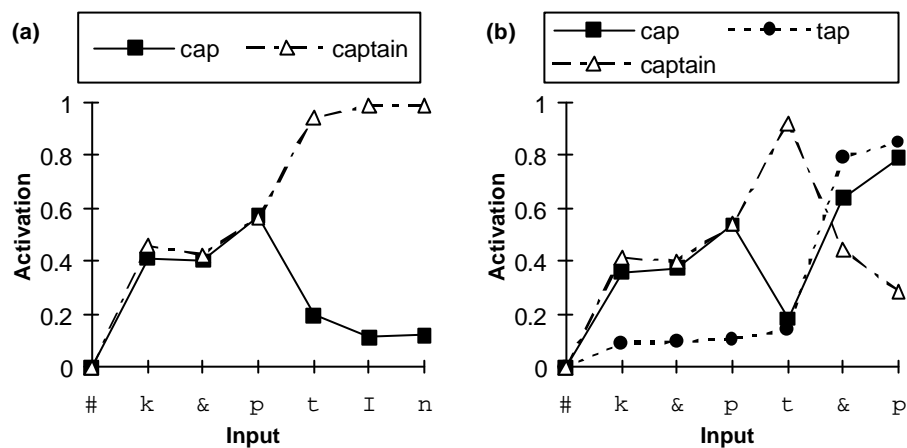

Figure 4: Activation of targets and competitors during sequences containing
(a) Bisyllables with embeddings (*captain/cap*) during *"captain"*
(b) Lexical 'garden paths' (*cap/captain*) during *"cap tap"*

2.2.2. Offset embedded words

The final set of results shown for these networks concern the identification of offset embedded words. By the account proposed in the original cohort theory [8] (where only words sharing the same onset are jointly activated) it would be predicted that the network would not activate offset-embedded words (e.g. *lock* in *padlock*) during recognition. Such a pattern of performance is illustrated in Figure 5. In contrast to the networks performance for onset-embedded words, the network is clearly capable of rejecting offset-embedded words during recognition.
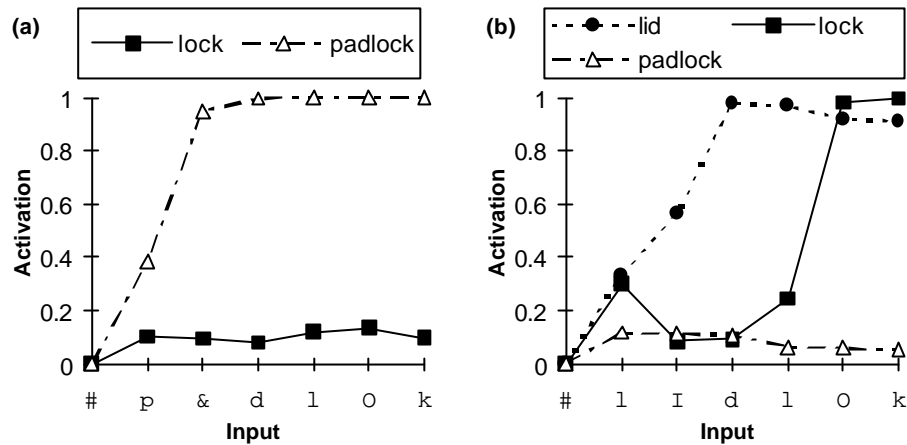
9

Figure 5: Activation of targets and competitors during sequences containing
(a) Bisyllables with offset embeddings (*padlock/lock*) during *"padlock"*
(b) Offset embedded words (*lock/padlock*) during *"lid lock"*

Empirical evidence on the activation of offset embedded words is unfortunately less clear. Several studies [26,27] report obtaining significant priming from offset-embedded words to an associatively related target (e.g. *trombone* primes RIB, via *bone*) though other experiments have failed to replicate this finding [28,29]. However, no results reported so far are inconsistent with the pattern shown in this network (and in models such as TRACE), where offset-embedded words receive substantially less activation that the longer words in which they are embedded.

3   Conclusions

The network described here learns to implement the maximal efficiency assumption in recognising words in sequences. Unlike previous accounts using simple recurrent networks [13,14], the network is able to deal with temporally ambiguous input not only where the ambiguity is resolved within a word, but also where post-offset information is required for recognition (as is the case for onset-embedded words). In recognising onset-embedded words the network displays an optimal compromise between partial activation of words for which the input is still ambiguous and full activation for unambiguous input. This holds even in cases where the network is presented with lexical garden-path sequences, where segments after the offset of an embedded word continue to match the longer competitor.

Furthermore, unlike previous simulations [18] the network is trained on input that does not contain any information about the location of word boundaries. We are therefore justified in claiming that the system is 'learning' lexical segmentation. At least for this limited training set, correspondences between form and meaning do provide a means by which a network could learn to identify individual words in

10

connected speech. Further simulations are required to investigate whether this method remains effective for more realistic vocabularies.

In making this claim we do not wish to suggest that this is the only means by which segmentation can be learnt. The results obtained by self-supervised and unsupervised methods [6,7] suggests that distributional analysis plays an important role in discovering lexical units in connected speech. Indeed, empirical evidence suggests that infants are capable of rapidly extracting distributional information from artificial speech *before* they have any knowledge of individual words [30]. However distributional analysis may operate more effectively in conjunction with mappings that identify meaningful units in connected speech. Further simulations, not reported here suggest that the architecture investigated here provides an ideal environment in which to explore the relationship between distributional analysis and lexical identification in the learning of segmentation. We show that adding a temporal auto-encoder mapping (an output trained to predict the next segment in the input) significantly speeds the network's acquisition of the recognition task. Further investigation of these networks' training profiles may therefore help to clarify the role of distributional analysis and pre-lexical segmentation cues in vocabulary acquisition.

3.1 Representational assumptions

The ability of these networks to recognise onset-embedded words is a direct consequence of using a target representation which preserves information about previous words in the speech stream. Instead of this representation changing with each word onset as in prior simulations [18], the target activation remains constant throughout each sequence of words. By using a representation that is static with respect to the speech stream, the network learns to segment speech into lexical units through extracting correspondences between form and meaning across different sequences. It is these correspondences that we consider to be the essence of lexical representation.

In constructing the output representations for these simulations we have used localist lexical units. Although the use of localist representations have considerable benefits in terms of ease of interpretation, they also introduce the unrealistic assumption that a word's lexical representation is categorical and invariant. Distributed output representations would allow the network to extract invariant representations from noisy and contextually variable meanings.

However, it is no straightforward matter to re-implement networks of this type using a distributed output representation. The useful property of localist lexical representations is that they provide a simple solution to the problem of how to activate multiple representations (of all the words in a sequence) whilst preserving similarity between the representation of the same lexical item in different positions. In a distributed scheme where lexical representations are non-orthogonal, multiple activation produces blend states with accompanying limitations on representational capacity [31]. To be able to use distributed representations we require some means of reducing the overlap between the lexical representation of different words, while at the same time preserving the similarity between the representation of the same lexical

item occurring in different positions. One such method would be to use a role/filler 'sentence gestalt' output [19] or a tensor product representation [32]. Such representations would also permit the incorporation of syntactic and constituent structures into these networks.

3.2 Comparing competing accounts

Having demonstrated the success of our model in learning to recognise onset-embedded words, it would be of interest to compare this model with previous accounts of spoken word recognition that incorporate lexical level competition [15,16]. This is made more difficult by the additional free parameters available in interactive activation models. However, one prediction made by both TRACE and Shortlist is that in recognising onset-embedded words greater competition between units representing long words (e.g. competition between *captain* and *captive*) would produce a short word advantage for input that was ambiguous between a short and a long word[*]. As can be seen in Figure 3, the networks investigated here produce lexical activations that approximate the conditional probabilities of all words that match the input, regardless of length. So for onset-embedded words such as *cap* and *captain* that occur with equal frequency in the training set, the network predicts equal activation of both candidates at the offset of *cap*.

Empirical evidence on this issue is inconclusive at present, particularly given the presence of acoustic cues such as vowel duration that distinguish the syllables of short and long words [25]. Indeed it may be that providing a psychologically realistic account of subjects sensitivity to duration differences in the speech stream presents a substantial challenge to accounts both with and without lexical competition. However the clear prediction made by recurrent network accounts (that responses to ambiguous input represent conditional probabilities) suggests that these models will reward future investigation. Further simulations are therefore required to investigate whether recurrent network models are able to account for the range of experimental data proposed as evidence for lexical competition.

Acknowledgements

---

[*] see for example Figure 3 on page 213 of Norris, 1994 [15]

in *Proceedings of the Fourth Neural Computation in Psychology Workshop*
J. Bullinaria, G. Houghton, D. Glasspool (eds). London: Springer-Verlag.
Final CRC - 1ˢᵗ August 1997

References

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

[2]   Nakatani, LH., Dukes, KD. Locus of segmental cues for word juncture. Journal of the Acoustical Society of America 1977; 62 (3): 715-719

[3]   Nakatani, LH., Schaffer, JA. Hearing words without words: Prosodic cues for word perception. Journal of the Acoustical Society of America 1978; 63 (1): 234-245

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

[5]   Cairns, P., Shillcock R., Chater N., Levy J. Bootstrapping word boundaries: a bottom-up corpus based approach to speech segmentation. In: Levy, J., Bairaktaris, J., Bullinaria, J., Cairns, P. (eds), Connectionist models of memory and language (289-310). UCL Press, London, (1995)

[6]   Christiansen, MH., Allen J., Seidenberg, MS. Learning to segment speech using multiple cues: a connectionist model. Language and Cognitive Processes (in press)

[7]   Brent, MR., Cartwright, TA. Distributional regularity and phonotactic constraints are useful for segmentation. Cognition 1996; 61: 93-125

[8]   Marslen-Wilson, WD., Welsh A., Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 1978; 10: 29-63

[9]   Luce, PA. A computational analysis of uniqueness points in auditory word recognition. Perception and Psychophysics 1986; 39: 155-158

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

[11] Bard, EG., Shillcock RC., Altmann GT. The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. Perception and Psychophysics 1988; 44: 395-408

[12] Elman, J. Finding structure in time. Cognitive Science 1990; 14: 179-211.

[13] Norris, D. A dynamic-net model of human speech recognition. In: Altmann, G. (ed) Cognitive Models of Speech Processing, MIT Press, Cambridge, MA, 1990

[14] Gaskell, MG., Marslen-Wilson, WD. Integrating form and meaning: A distributed model of speech perception. Language and Cognitive Processes, in press

[15] McClelland, JL., Elman, JL. The TRACE model of speech perception. Cognitive Psychology 1986; 18: 1-86

[16] Norris, D. Shortlist: a connectionist model of continuous speech recognition. Cognition 1994; 52: 189-234

[17] Norris, D., McQueen JM., Cutler, A., Butterfield, S. The possible-word constraint in the segmentation of continuous speech. (submitted).

[18] Content, A., Sternon, P. Modelling retroactive context effects in spoken word recognition with a simple recurrent network. In: Ram, A., Eiselt, K. (eds)

Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, Lawrence Erlbaum, Hillsdale, NJ, 1994

[19] St. John, MF., McClelland, JL. Learning and applying contextual constraints in sentence comprehension. Artificial Intelligence 1990; 46: 217-257

[20] Plunkett, K., Sinha C., Møller MF., Strandsby, O. Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. Connection Science 1992; 4 (3/4): 293-312

[21] de Sa, VR. Unsupervised classification learning from cross-modal environmental structure. PhD Thesis, University of Rochester, New York, 1994

[22] Ladefoged, P. A course in phonetics. Second Edition Harcourt Bruce, New York, 1982

[23] Servan-Schreiber, DA. Cleeremans A., McClelland, JL. Graded-state machines: The representation of temporal contingencies in simple recurrent networks. Machine Leaning 1991; 7: 161-193

[24] Gaskell, MG., Marslen-Wilson, WD. Measuring lexical competition in speech perception. submitted

[25] Davis, MH., Marslen-Wilson, WD., Gaskell, MG. Ambiguity and competition in lexical segmentation. In: Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society (in press)

[26] Shillcock, RC. Lexical hypotheses in continuous speech. In: Altmann, G. Cognitive Models of Speech Processing, MIT Press, Cambridge, MA, 1990

[27] Vroomen, J., de Gelder, B. Activation of embedded words in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance 1997; 23(3): 710-720

[28] Marslen-Wilson, WD., Tyler LK., Waksler R., Older L. Morphology and meaning in the English mental lexicon. Psychological Review 1994; 101(1): 3-33

[29] Gow DJ., Gordon PC. Lexical and pre-lexical influences on word segmentation. Journal of Experimental Psychology: Human Perception and Performance 1995; 21(2): 344-359

[30] Saffran, JR., Aslin, RN., Newport, EL. Statistical language learning by 8 month-old infants. Science 1996; 274(5294): 1926-1928

[31] Gaskell, MG. Parallel activation of distributed concepts: Who put the P in the PDP? In: Cottrell, GW. (ed) Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society, Lawrence Erlbaum, Mahwah, NJ, 1996

[32] Smolensky, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial Intelligence 1990; 46: 159-216