6. Acoustic cues to word length in recurrent networks

Simulations described in Chapter 3 demonstrated that a recurrent network trained to activate a representation of all the words in a sequence provides an account of the delayed recognition of onset-embedded words. Following training, the network activates all lexical units that match the current input, with the degree of activation representing the probability of each word given the current input. Where the input matches both a complete word and the start of a longer word the network will activate each item equally. Only where following context rules out the longer lexical item will the network fully activate the embedded word.

This behavioural profile (illustrated in Figure 3.4 and 3.5) contrasts with that predicted by models that incorporate lexical-level competition. Lexical competition models such as TRACE (McClelland & Elman, 1986) produce a bias towards short word interpretations at the offset of an embedded word. This is due to the greater number of competitors that are present for longer words. This short word bias benefits embedded words since it allows them to inhibit longer competitors and thus be recognised more easily.

As was described in the previous chapters, both of these behavioural profiles are inadequate as an account of the time course with which embedded words are identified. The results of gating and cross-modal priming experiments demonstrate that short and long words can be distinguished before the offset of the first syllable of the embedded word. Therefore, since neither the lexical competition models nor the recurrent network simulations in Chapter 3 incorporate any input cue that would serve to distinguish between short and long words, both accounts are at present insufficient to account for the experimental data presented in Chapters 4 and 5.

An important goal of the modelling reported in this chapter is to produce a computational account of these experimental data. At present, recurrent network accounts of spoken word recognition are incomplete with respect to the processing of cues to word length and word boundaries. Simulating the effect of these cues on the identification of embedded words not only allows evaluation of whether recurrent networks are sufficient as an account of the processing of onset-embedded words, but will also allow testable predictions for future experiments to be generated from the model.

6.1. Acoustic cues to word length

The current hypothesis was that differences in segment and syllable duration provide the acoustic cue to word boundaries required to account for the experimental data. As described in the review of the acoustic-phonetics literature presented in Chapter 2, the increased duration of syllables in monosyllabic words has been reliably reported (Klatt, 1976; Lehiste, 1972) and there is experimental data supporting the use of duration as a cue to the perception of word boundaries (Nakatani & Schaffer, 1978). Furthermore, as shown in Table 4.1 in Chapter 4, differences in duration between syllables such as /kæp/ in words like *cap* and *captain* are present in the stimuli used in the current series of experiments.

Sensitivity to differences in syllable duration would therefore provide an account of listeners' ability to discriminate between short and long words before the onset of the following word – as shown by differences between responses to short and long word stimuli at AP_1 both in gating (Experiment 1) and in cross-modal priming (Experiment 2a). Thus, although the critical experiments directly manipulating duration have yet to be carried out, current evidence suggests that the discrimination of short and long words may involve the detection of differences in segment and syllable duration.

6.1.1. The perception of duration differences

An important constraint in modelling the perception of changes to segment and syllable duration in the recognition of onset-embedded words is that duration can not be used as a deterministic cue to whether a word is monosyllabic or bisyllabic. Despite the significant differences between the duration of the syllables in short and long word stimuli ($p<.001$ by a paired t-test), as shown by the histogram in Figure 6.1 there is considerable overlap in the distribution of durations for syllables in short and long words.

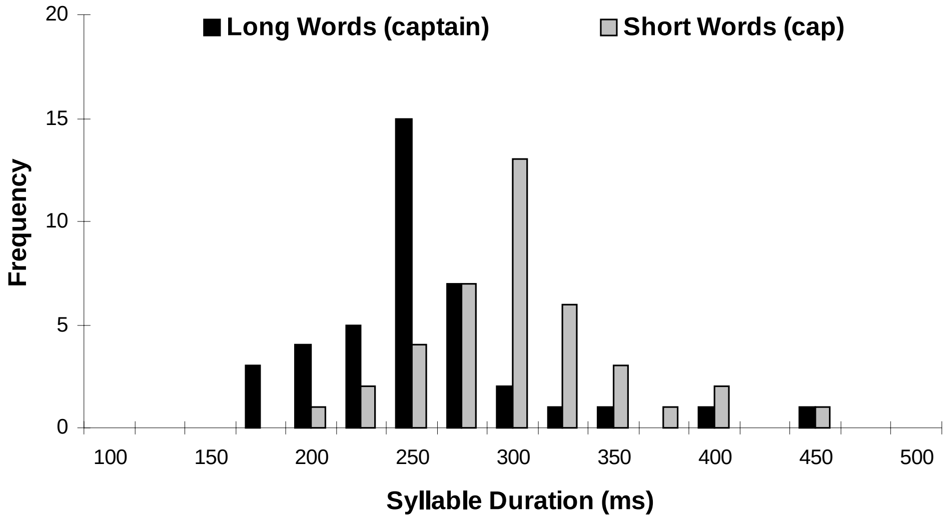

Figure 6.1: Histogram of the duration of target syllables in short and long test items

Such overlap will be even more marked in comparisons of naturally occurring speech. Significant differences in syllable duration in short and long words are only found where additional sources of variance can be controlled for. These may be caused by measuring syllables produced by different speakers, at different positions in an utterance, and containing different constituent segments and associated stress (Klatt, 1976). Without compensating for these additional sources of variance it will not be possible to use syllable duration to distinguish short from long words (see for instance the discussion between Crystal and House (1990) and Anderson and Port (1994) regarding the reliability of duration as a cue to word boundaries in English).

From a computational perspective, this implies that it will not be possible to partition syllables into those coming from short and long words by setting a simple duration threshold. Instead, some additional process will be required to adapt to the ongoing speech stream so that information from the spoken context can be used to set an appropriate boundary for distinguishing syllables in short and long words. This process is required to allow the system to compensate for other sources of variation in syllable duration in order to identify syllables as coming from a short or a long word.

In modelling this adaptive process, the simulations reported in this chapter focus on an extreme form of ambiguity – the case where two syllables with identical durations in fact come from words of different length. Syllables such as these can still be used to distinguish between short and long words if they are produced in spoken contexts that lead the listener to expect a syllable of a particular duration. For instance, if a sequence is produced at a fast speech rate, a syllable from a monosyllabic word would be produced with a relatively short duration. A syllable with an identical duration but produced in a slower sentence, on the other hand, would be more likely to have come from a bisyllabic word.

Such cases – embedded syllables in short and long words occurring with the same duration – only arose for a small number of the experimental stimuli. This is due to the test syllables being controlled for many potential sources of variation in syllable duration. These tightly controlled stimuli therefore provided less sources of variation that could lead to syllables in short and long words being produced with identical durations. However, for more naturally produced speech, in which these sources of variation are uncontrolled it is likely that the duration of syllables in short and long words will need to be disambiguated by the contexts in which they are produced.

In simulating the processing of duration as a cue to the placement of word boundaries, the models developed in this chapter used, as a test case, stimuli in which identical durations were produced for syllables in short and long words. This will allow investigation of the processes by which compensation for contextual changes in syllable duration are applied during recognition. These simulations focus on just one of the possible variables that can affect duration in such a way as to make syllables ambiguous – a change in the overall rate at which speech is produced in sequences.

6.2. Simulation 3 – Processing syllable duration in fast and slow sequences

The adaptive processing of acoustic input is an important aspect of the perception of connected speech. A system that can compensate for differences in the acoustic properties of speech produced by different speakers at different rates is also required to account for the perception of time-compressed speech (Foulke & Sticht, 1969; Dupoux & Green, 1997; Pallier, Sebastian-Galles, Dupoux, Christophe, & Mehler, 1998). The simulations reported here will investigate whether the recurrent network architecture that was described in Chapter 3 is able to adaptively process an input cue analogous to changes in duration in different rate sequences. This will require the network to process identical words differently, depending on the rate at which the preceding words in an utterance are presented.

6.2.1. Representing duration information in connectionist networks

In order to investigate the processing of duration in recurrent networks a decision must be made about how duration is to be represented. Since these simulations use an artificial language coded as discrete segments there is considerable freedom to represent duration in a form that makes the network's task as easy as possible. However, it is also important that the representation should not make unrealistic assumptions about the information available in the speech stream.

Previous computational models have coded for duration information by duplicating segments over many time steps in the input (Abu-Bakar & Chater, 1995; Gupta & Mozer, 1993). This may incorporate the unrealistic assumption that longer segments will have identical spectral properties to shorter segments but extended over more time steps. This simplifying assumption may help processing in the network since it creates greater similarity between identical sequences presented at different rates. However, coding for duration in this way is also computationally expensive since it requires extended training and testing sets to include these duplicated segments. Furthermore, in order to incorporate effects of preceding context, the network must encode information over more intervening segments. It was consequently decided to use a more computationally tractable coding scheme for segment and syllable durations.

The method of coding for duration that is most tractable for the network is to use input units that (separate from all other inputs) provide information about the duration of the current segment or syllable. This representation assumes that duration can be coded in an equivalent way to any other aspect of the speech input. Indeed, from the networks' point of view, these additional units could represent any source of information that helps distinguish between short and long words – not just differences in duration. Since these network investigations are intended to simulate the results of Experiment 2 (where duration differences were only one of several possible acoustic cues that could distinguish between short and long words) such a representation may be easier to justify in this context. In the remains of this chapter, however, the input representation will be set using an assumption that these units are coding for duration. This will allow us to make clear predictions regarding the computational mechanisms that may be required in simulating the affect of duration cues on lexical activation.

The question then becomes how to represent duration at this input. Exploratory simulations added a single input unit which represented duration by its activation (i.e. high activation for long syllables, low activation for short syllables). However, these simulations proved unsuccessful since the small changes in activation produced in this single unit were swamped by the binary inputs representing phonetic features.

Consequently a binary representation of duration over a block of three units was used in the network. These provided for three duration codes along with an additional input unit that was active whenever duration information was inappropriate for the current input segment (i.e. the unit was set to zero in the gaps between sequences and in certain syllable positions in later simulations). These codes allow a context-dependent representation of the duration of syllables to be used. This provides a simple simulation of the overlapping distributions of syllable duration that may result from changes in the overall rate at which a sequence is generated.

Syllable			**Speech Rate**	
Duration	**Code**	**Network Input**	**Fast**	**Slow**
no duration	0	0 0 0	-	-
short	1	1 1 1	bisyllable	-
medium	2	0 1 1	monosyllable	bisyllable
long	3	0 0 1	-	monosyllable

Table 6.1: Network representation of syllable duration in Simulations 3 and 4.

The activation of the three input units in coding for three values of duration (plus the ‘no duration’ input) is shown in Table 6.1. As can be seen, the longest syllable duration (code 3) is only used for monosyllabic words, while the shortest duration (code 1) is only used to represent bisyllables. However, the intermediate duration (code 2) can be used for either a monosyllabic or a bisyllabic word depending on the rate at which the sequence is being produced. Thus, for a sequence produced at a ‘slow rate’, the longest duration (code

3) will be reserved for monosyllables and code 2 will represent bisyllables. For sequences produced at the ‘fast rate’ code 2 will represent monosyllables and code 1 will be used to represent the duration of syllables in bisyllabic words. Thus code 2 will be ambiguous; depending on the overall ‘rate’ at which the sequence is presented; syllables coded with this duration can either come from short (monosyllabic) or long (bisyllabic) words.

For each sequence of words generated in the training set a speech rate was selected at random. This determined which two of the three duration codes would be presented to the network for that sequence of words. Since syllables coded as duration 2 are ambiguous, the networks will need to use the duration associated with previous words in the current sequence to determine the overall speech rate for that sequence. For onset-embedded words, where the identity of segments does not distinguish between short and long words, the network should be able to use duration to contribute to the identification of a word with an ambiguous syllable (such as /kæp/ in *cap* and *captain*). The network’s sensitivity to the duration cue was tested by comparing the activation of an onset-embedded word occurring at the start of a sequence (where duration code 2 will be ambiguous and will not provide any cue to discriminate short from long words) with the same input occurring as the second word of a sequence (where prior context should allow the network to determine the overall ‘rate’ of that sequence, and disambiguate the duration code).

6.2.2. Network architecture and training set

The architecture used for the initial simulations was identical to that used in Simulation 2 reported in Chapter 3 (see Figure 3.7) except for the addition of 3 input units and prediction outputs to represent the duration codes shown in Table 6.1. These simulations therefore used a simple recurrent network with 9 input units, 50 hidden units copied back to 50 context units and 29 output units. As in the simulations reported previously, the lexical output units had no bias weights although these were retained for the auto-encoder outputs. The architecture of the network and a snapshot during training is shown in Figure 6.2.

Training sets for these simulations were constructed using the 20 word vocabulary shown in Table 3.2 coded using the distributed phonetic feature representation shown in Table 3.1. One difference between this and previous simulations was that for each sequence of between 2 and 4 words one of the two speech rates described in the previous section was chosen at random. This determines which two of the three duration values are used to code for short and long words in that sequence. In the initial set of simulations, all segments in all syllables were coded for duration. This training regime is illustrated in Figure 6.2 where the network is being trained on the sequence "*lid topknot*" at the fast rate. The monosyllable *lid* is therefore presented with a duration of 2 while the prediction output for the first segment of the bisyllable *topknot* is being trained to predict a syllable with duration 1.

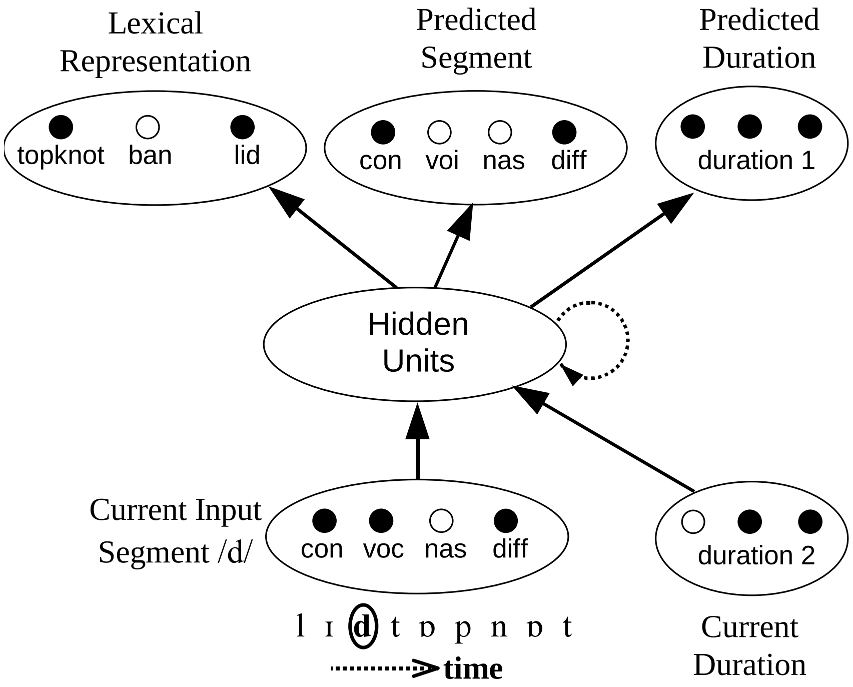

Figure 6.2: A snapshot of the SRN during training for the sequence "*lid topknot*" at the fast rate in simulation 3. Three additional input units and prediction units represent the duration of the current syllable and the predicted duration of the subsequent segment.

Preliminary simulations showed that the network was too reliant on duration information. For instance, since there was only one bisyllabic word (*topknot*) that began with the segment /t/, for sequences where the unambiguous duration code was used (code 1), the effective uniqueness point for the network was at the initial segment of the word. This inappropriate performance is apparent because of the small number of vocabulary items in the network's training set. However, this unrealistic aspect of the simulation is caused by the network using duration to rule out segmentally appropriate hypotheses. This suggests that the duration code is too reliable to simulate the properties of the speech stream, reflecting the discrepancy between simulations in which only one variable affects syllable duration and the more realistic case in which multiple, unreliable variables interact to determine the duration of segments and syllables in the input. In order to approximate the statistical properties of connected speech more closely, the duration cue was made less reliable in subsequent simulations. This was achieved by replacing 20% of the words coded with unambiguous durations (codes 1 and 3) with the ambiguous duration (code 2).

6.2.3. Results

Ten networks were trained using different random seeds for generating the training sequences and different sets of random initial weights. Lexical activations for onset-embedded words and competitors (both presented with the ambiguous duration code 2) were recorded and averaged over 10 fully trained networks. Results shown in Figure 6.3 compare the activation of onset-embedded words and longer competitors for sequences where the initial syllable of both items is coded as duration 2 (and could therefore come from either a short embedded word or a longer competitor).

The results shown in Figure 6.3a and Figure 6.3c essentially replicate the pattern of performance shown in simulations without duration information. Without preceding context the network is unable to identify the underlying speech rate for these sequences. Consequently it processes the input segments representing the embedded words identically – regardless of whether the embedded syllable comes from a short or a long word. In these circumstances the network is forced to use following context to resolve this ambiguity. Since these networks receive less training on inputs requiring the use of following context than the simulations reported in Chapter 3, their performance is marginally worse than the networks trained without duration cues in Simulation 1. However, as can be seen by comparing these graphs with Figure 3.4b and Figure 3.5a, the activation profile of both sets of networks is qualitatively similar.

By comparison, Figure 6.3b and 6.3d show that the network where preceding contexts are available, the network can use the duration of the preceding word to determine the rate at which the current sequence is being produced. Consequently, although the input for the syllable /kæp/ is identical in both cases (duration code 2), these networks can determine whether the ambiguous duration is more likely to have come from a short or a long word and increase the activation of the appropriate lexical unit accordingly. Both *cap* and *captain* are presented with a single word of preceding context, with an unambiguous value of duration in each case. This difference in duration allows the networks to determine whether the target syllable is faster or slower than the unambiguous preceding word allowing them to process an otherwise ambiguous input correctly. These networks therefore show increased activation of the appropriate lexical units before the offset of the embedded word.

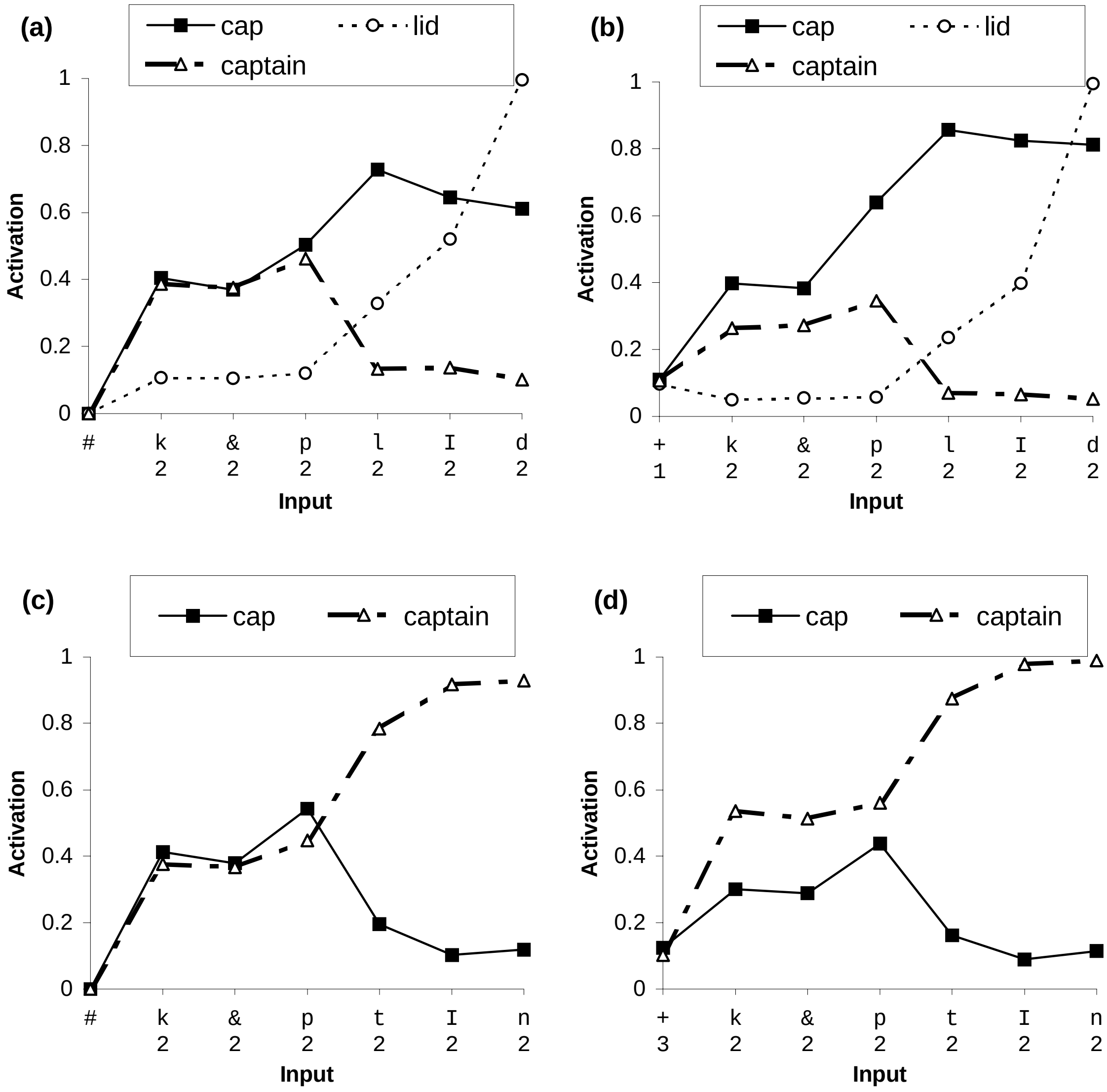

Figure 6.3: Activation of an onset-embedded word (*cap*) and competitor (*captain*) averaged over ten networks in Simulation 3. Input segments are presented with the duration codes shown in Table 6.1. Embedded words occur either as the first word in a sequence (Figures a and c, preceded by a sequence boundary marked #) or as the second word in a sequence (Figures b and d, preceded by another word marked + with an associated syllable duration).

Example sequences are:

(a) *"cap lid"*	(fast rate)	**(b)** bisyllable + ***"cap lid"***	(fast rate)
(c) *"captain"*	(slow rate)	**(d)** monosyllable + ***"captain"***	(slow rate)

This difference between the activation profile of onset-embedded words at the start or in the middle of a sequence, makes the prediction that if syllable durations overlap in short and long words, then the ambiguity created by onset-embedded words will be greater when embedded words are presented in isolation or at the onset of a sentence. Without information from prior context by which to determine speech rate it may not be possible for the recognition system to account for contextual variation in order to use syllable duration as a cue to the length of the target word. Since the majority of experiments looking at word identification have used single word stimuli this may explain the absence of duration effects in the literature reviewed in Chapter 4. Further experimental work should aim to investigate whether – as observed in this simulation – differences in the duration of preceding syllables can alter the perception of an otherwise ambiguous embedded word (as illustrated in Figure 6.3b and d).

However, one aspect of this model may be inappropriate as a simulation of the processing of duration cues to word length. Since duration information is presented at all positions in a syllable it is possible for the model to use changes in the duration of adjacent segments as a cue to the location of a word boundary. Given the importance of transitional information in the processing of simple recurrent networks (Elman, 1990; Servan-Schreiber, Cleeremans, & McClelland, 1991) it is likely that this information plays a role in the network's use of duration as a cue to word length. This transitional information constitutes an unrealistic aspect of the model since these transitions carry information that would not be present in connected speech where only certain speech segments may change with syllable duration.

To take a concrete example, both of the input sequences shown in Figure 6.3b and Figure 6.3d used an unambiguous duration prior to the embedded word. Consequently, there was a change in duration at the onset of the target word that was distinctive to sequences in which the target word was a short word (Figure 6.3b) or a long word (Figure 6.3d). It is possible that this change in duration (rather than duration per se) provides a cue to the length of the target word.

By comparison the test sequences shown in Figure 6.4 illustrate the processing of sequences where two successive words have an ambiguous duration value. Where both words are presented with duration code 2, there will be no change in duration at the transition between words to provide a cue to the length of the target word. Instead, the network must use the duration of the first word in combination with its lexical identity to determine the rate at which the sequence is being presented.

The most common case in the networks' training sets where two successive words are presented with an ambiguous duration is where two monosyllables are presented in a sequence at the fast rate. The networks' processing of this input is illustrated in Figure 6.4a. As can be seen the network still succeeds in using the duration of the word to increase the activation of the monosyllable over its longer competitor before the acoustic offset of the word.

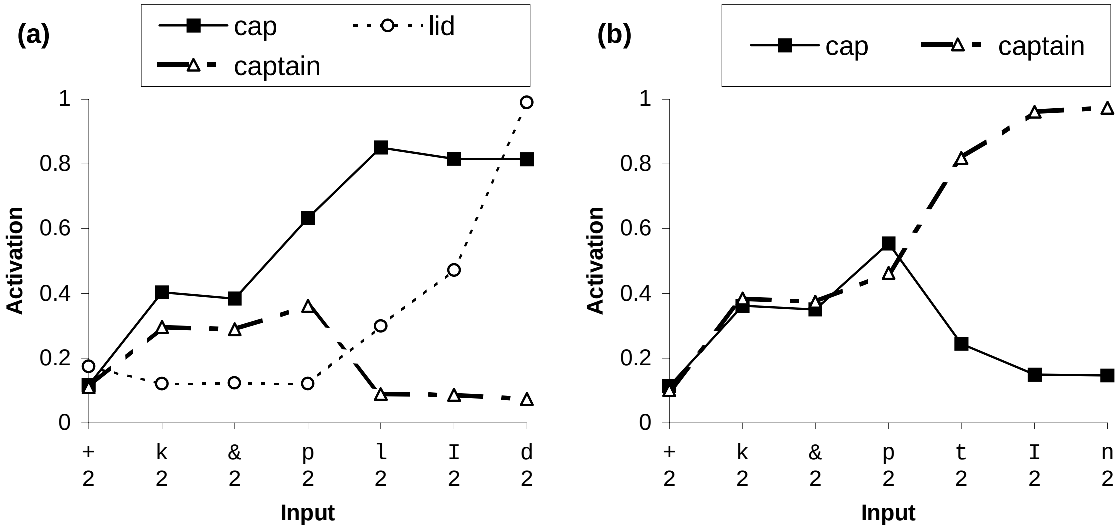

Figure 6.4: Activation of an onset-embedded word (*cap*) and competitor (*captain*) in Simulation 3. All syllables in these test sequences were presented with an ambiguous duration.

Example sequences are:
(a) monosyllable + ***"cap lid"*** (fast rate) **(b)** bisyllable + ***"captain"*** (slow rate)

A similar pattern of two successive ambiguous duration codes can also occur where two bisyllabic words are presented consecutively in a sequence at the slow rate. This case will occur less often in the training set (since there are fewer bisyllabic words than monosyllabic words in the language). The network's processing of this input is illustrated in Figure 6.4b. As can be seen in this graph the networks no longer favour the appropriate lexical item at the offset of the embedded word. Comparing Figure 6.4b with Figure 6.3d shows that the networks' activation profile in identifying bisyllables containing an embedded word is only altered by the duration of the previous word where there is a change in duration at the word boundary. If the lexical identity of the previous word as well as its duration must be used to establish the rate at which the sequence is presented (where the duration associated with both words in a sequence are ambiguous) the networks is not always able to use prior context in processing ambiguous input.

6.2.4. Discussion

These networks are clearly able to process input that includes a representation analogous to duration. Although not shown here, the network's responses to unambiguous duration codes are clear and categorical. However, as was described in the conclusions of Chapter 5, because of the amount of variation in segment and syllable duration (both between and within speakers) it is likely that duration will seldom provide an unambiguous or absolute cue to the number of syllables in a word. Consequently, in modelling sensitivity to duration cues, this simulations has focused on a test case in which the input to the network would be ambiguous without prior context (i.e. it could come from either a monosyllable or a bisyllable). In this case embedded words can only be disambiguated where prior context is used to detect the underlying 'rate' of the sequence. Since the overall rate at which a sentence is produced has been shown to effect voice onset time (VOT) boundaries for the perception of stop consonants (Wayland, Miller and Volaitis, 1994), it might be expected that equivalent results would be observed in the perception of monosyllabic and bisyllabic words. However, further experiments are required to demonstrate that the same adaptive properties are observed in the use of syllable duration as a cue to word length and word boundaries.

As discussed in conjunction with Figure 6.3 there is clear evidence that the network is able to use duration as a cue (relative to preceding context) to determine whether an embedded word is more likely to have come from a short word or a long word. However detailed investigation suggests that the networks' use of this information is more efficient for stimuli in which the previous word has an unambiguous duration. This suggests that it is the transition between different duration syllables that carries the most salient information for the network. It is therefore necessary to establish that the network can make use of duration information where a more appropriate representation of the speech stream is provided.

6.3. Simulation 4 – Representing duration from vowel to vowel

It has been shown in Simulation 3 that the networks investigated here are capable of using inputs corresponding to duration as a relative cue. However the input representation used for this network provides a more salient duration cue than might be found in the speech stream. Most importantly, the input representation assumes that duration information is present in all of the segments that make up a word. Although prior work has shown that changes in speech rate will manifest themselves across an entire syllable (for example VOT will change for a word-initial stop consonant depending on overall syllable length, Miller (1979)), it is unlikely that these differences in rate could be detected from the onset of a syllable. Indeed there is a discrepancy between the position at which changes arise as a result of altered syllable durations and the point at which syllable duration can be detected in order to compensate for these changes. This has been a focus of recent modelling work on the effects of speech rate on phonetic categorisation (Abu-Bakar & Chater, 1995).

Effects of syllable duration on phonetic categorisation suggest that the overall duration of a syllable can not be established until late on during its presentation. Consequently transitional information that was responsible for the processing of duration cues in Simulation 3 is unlikely to be available in real speech. A more realistic assumption would be that duration information is only available late on during the processing of a syllable. Since the majority of variation in syllable duration is caused by changes in the vowel (Klatt, 1976) a further set of simulations were therefore carried out in which duration input is only presented for vowel segments.

6.3.1. Network architecture and training set

The network architecture, input and output representations and vocabulary remained the same for this simulation as in Simulation 3. The only change made was that the duration input was only activated for the vowel of each syllable. By providing duration information in vowels only, the network is no longer able to detect changes in syllable duration at word boundaries. To make use of prior context in processing ambiguous values of duration, the network will have to retain information about the preceding syllable across at least two intervening, unmarked segments. In order not to penalise the network for continuing to activate the duration output after the vowel segment, no error was propagated back from the duration units at the prediction output unless the predicted segment was a vowel.

Results of initial simulations carried out using this training regime were disappointing – the network showed no ability to use prior context in processing syllables with ambiguous durations. Graphs equivalent to those in Figure 6.3 showed that the pattern of activation in these networks was identical, irrespective of whether ambiguous (embedded) syllables occurred at the start or as the second word of a sequence. This finding confirms that in Simulation 3 it is transitional information between syllables that enable the networks reported previously to display sensitivity to prior context in processing the duration information. Where duration information is separated by unmarked segments, networks were unable to use prior context in processing the ambiguous input.

These results are reminiscent of those reported by Elman (1993) showing that it is difficult to train SRNs to process long distance dependencies where unrelated information is presented at intervening time steps. This is a consequence of the SRN having limited access to representations of states at previous time steps. Unless relevant aspects of the prior input are represented in the hidden units at the time step before it needs to be used then the network will be blind to prior context. Since two unmarked time steps separated the duration inputs for successive syllables, these networks did not retain a representation of the duration of the previous syllable that would allow them to use prior context.

One solution that has been used in training networks faced with this problem is to use fully recurrent networks in which error is propagated back over copies of the system’s internal representations extending over more than one time-step (Rumelhart, Hinton, & Williams, 1986). Such an approach is not without its problems, however, since information represented several time steps previously will have a decreasing influence on the error gradients that drive the learning algorithm. Another approach used to encourage the network to retain an internal representation of the necessary information was described by Maskara and Noetzel (1993). This requires the network to output a copy of the hidden unit activations from the previous time step. Such an approach is reported to be successful in learning centre-embedded sequences that SRNs find difficult.

The approach taken in this thesis was to extend the prediction task that was used previously. As in simulations reported by Shillcock, Levy and Chater (1991) and Gaskell, Hare and Marslen-Wilson (1995) output units were added that not only represented the following segment and duration to be presented in the input, but also the current and previous input. In order to activate a representation of the identity and duration of segments presented at preceding time steps the network must retain an appropriate representation at the hidden units. This internal representation helps ensure that the network has access to segment information and syllable duration from the previous word – assisting the learning of the duration cue.

Aside from the additional output units for this extended output task, all other aspects of the network were identical to those reported previously. A simple recurrent network was used with 9 inputs, 50 hidden units copied back to 50 context units, and 47 output units – 20 lexical units and 9 output units (6 for phonetic features, 3 for duration) for each of the 3 prediction outputs. The network was trained on the same vocabulary used before with duration represented over three units using the coding scheme shown in Table 6.1. However, in this simulation, duration information was only presented at the input for the vowel of each syllable. Similarly, error was only propagated back from output units representing the duration of vowel segments.

6.3.2. Results and discussion

Ten networks were trained for 500 000 sequences using this architecture and training regime. All results reported below are the average of ten training runs, each having different initial weights and randomly generated training sequences. In contrast to preliminary simulations that did not include output units representing the current and previous input, these networks were successful in utilising prior context to identify the onset-embedded words in the training set. In all four combinations of preceding word and ambiguous target syllable shown in Figure 6.5, the model produces increased activation of the correct lexical item at the offset of the syllable forming an embedded word.

Therefore, incorporating additional output units representing the current and previous input segments enabled these networks to retain duration information from the vowel of one syllable to the vowel of the next. By retaining this additional information over several intervening time steps the network was able to use duration cues in cases where the previous word, as well as the current word, is presented with an ambiguous syllable duration – unlike the networks described in Simulation 3.

The addition of output units that encourage the network to retain a more complete representation of the prior context appears to increase the networks ability to retain lexical information about the preceding word. This allows the system to use not only the duration of the preceding word but also its lexical identity to detect the rate at which a sequence is presented. Hence these networks can process ambiguous target syllables more effectively than the networks in Simulation 3.

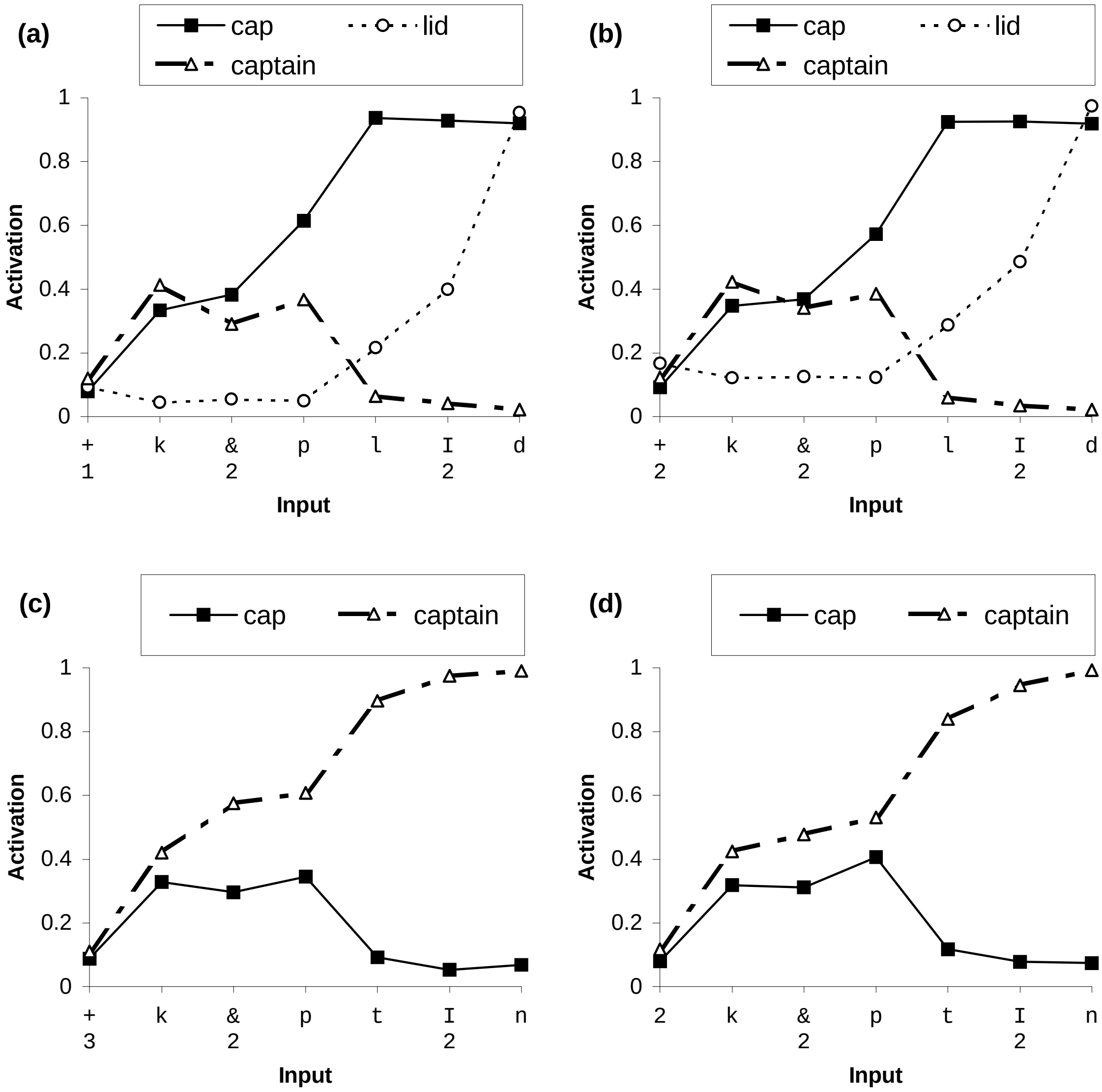

Figure 6.5: Activation of onset-embedded words (*cap*) and competitors (*captain*) with ambiguous durations in Simulation 4. Target syllables occur as the second word of a sequence, preceded either by a word with an unambiguous duration (a and c) or an ambiguous duration (b and d)

Example sequences are:

(a) bisyllable + "***cap lid***" (fast rate) **(b)** monosyllable + "***cap lid***" (fast rate)
(c) monosyllable + "***captain***" (slow rate) **(d)** bisyllable + "***captain***" (slow rate)

It is of interest that this property was only observed where additional output tasks were used to force the network to represent previous input at the hidden layer. In all the simulations reported so far the network must still activate a representation of the identity of the preceding word at the lexical units. However, it appears that in order for the networks to use the identity of previous words this information needs to be represented in a particular form at the hidden units – merely activating the appropriate lexical unit is insufficient. These additional output tasks therefore play a valuable role in structuring the networks' internal representations to enable them to use duration information in identifying lexical items. A similar argument was also applied to the faster learning observed in Simulation 2 reported in Chapter 3; comparing networks with and without the prediction task suggests that the additional task helps structure the networks' internal representations of the input to assist in learning the lexical identification task.

The networks in Simulation 4 are therefore capable of using duration information to disambiguate onset-embedded words in sequences in which the detection of durations associated with short and long words must be adjusted by consideration of the rate at which preceding words in a sequence were presented. Adaptive processing such as this is likely to be required in order to use the duration of syllables in connected speech as a cue to the location of word boundaries. It therefore seems appropriate to compare the activation profile observed in the networks reported in Simulation 4 to the results of the experiments reported in the previous chapters.

6.4. Simulating experimental data

Given concerns over the role of response biases in gating, cross-modal priming is likely to provide a more transparent measure of lexical activation than gating data. A further assumption in simulating priming data is that the magnitude of priming is directly proportional to the activation of the relevant lexical output unit in the network. However, since priming effects are not equivalent to lexical activations in a network it is inappropriate to transform the activation scale into priming units. Comparisons between experiments and simulations will be facilitated, however, by plotting priming and simulation data side-by-side and by using the same scale in all four sets of experimental data and simulation results.

6.4.1. Method

An important goal in simulating the priming data will be to ensure that statistical analyses of the networks' activations produce the same results as those obtained for behavioural data. Since there are only two onset embedded words and two longer competitors in the network's vocabulary it will not be possible to do analyses over different items. Consequently, statistical analysis will focus on analysing whether a pattern of results is reliable across all ten networks trained, treating each network as a single subject. Experimental evidence suggested that there were two sources of information that are relevant to the identification of onset-embedded words and longer competitors. We will describe how these sources of information were simulated in turn.

Duration cues to word boundaries

The first source of information involved in the identification of onset-embedded words is the acoustic difference between syllables of monosyllabic and bi-syllabic words. Simulations reported in this chapter represent this acoustic cue as a difference in the activation of a group of units that code for syllable duration. As discussed previously, this may be an unrealistic representation of the speech stream. However, this simulation captures the adaptive nature of the acoustic processing of duration cues to word boundaries: specifically that sensitivity to the duration cue requires a comparison of the current syllable with that of preceding words in the sequence. In order to incorporate this property into the simulation, the critical syllables of the test stimuli were presented with an ambiguous duration (i.e. both monosyllabic and bisyllabic words used duration code 2 from Table 6.1). Since the stimuli used in the priming experiments have several syllables of preceding context, the test sequences for the network were presented with one word of preceding context to allow the duration input to be disambiguated.

The test sequences contained a mixture of preceding contexts with both monosyllabic and bisyllabic words. Preceding contexts for the onset-embedded words will exclude the longer words in which they are embedded and vice-versa. In all cases, however, the target syllable will be of an ambiguous duration (code 2). These target words are more appropriate in simulating the results of the priming experiments since prior duration is required to process the duration of target syllables in these test stimuli correctly.

Continuations of short word stimuli

The second source of information that plays a role in the identification of onset-embedded words and longer competitors is the segment that follows the offset of the embedded word. In our test stimuli this continuation matched a longer word. Priming results for these lexical garden-path stimuli suggested that longer lexical items continued to be activated. It is only at later probe positions where there is mismatch between these continuations of short word stimuli and longer lexical items that significant priming of embedded words in the absence of priming of longer competitors was obtained. Thus it was argued that lexical garden-path stimuli – such as the sequence *cap tucked* – played an important role in producing the activation profile shown for short word stimuli in cross-modal priming.

To simulate the results of these priming experiments, embedded words were therefore placed in lexical garden-path contexts equivalent to those used in the experimental stimuli. These lexical garden paths were generated such that a monosyllabic word matching the onset of the longer lexical item followed the embedded word (i.e. sequences used were of the form *cap tap* rather than *cap topknot*). This will ensure that changes in the duration of subsequent input do not provide an additional cue that following segments come from a separate word.

Probe positions and alignment points

In comparing the time course of activation of short and long words in these sequences, it is important to ensure not only that the stimuli are appropriate for comparison with experimental data, but also that the probe positions match those that were used in the experiments. Since the input to the network is divided into discrete segments, information in the input can be specified more precisely than was possible for the speech used in the experiments. However, in the initial comparisons of experiment and model, it is simplest to assume that the information assumed to be available at each alignment point in the experimental stimuli is available to the network.

It is assumed that the stimuli up to AP_1 contain information about the identity of the first syllable but no information about following segments. The section of speech between AP_1 and AP_2 was assumed to contain information relating to the identity of the onset of the following word, but no information that mismatches with the longer lexical item. After

AP_2, the stimuli containing short and long words diverge phonemically as information in the vowel (up to AP_3) and offset (up to AP_4) of the second syllable is presented to the network.

Thus, for one of the two pairs of test sequences in the network (c*aptain* and *cap tap*, rather than *bandit* and *ban dock*), AP_1 was following the segments /kæp/ with the vowel being presented with duration 2. Thus, this syllable will be identical for stimuli containing short and long words. However, it is expected that the network will be able to use prior context in disambiguating this input. As was intended for the experimental stimuli, AP_2 includes the initial segment of the following word e.g. /kæpt/. It is only at AP_3, following the input /kæptɪ/ for *captain* and /kæptæ/ for *cap tap* that test stimuli will diverge phonemically. The final probe position AP_4 is assumed to be at the offset of both words /kæptɪn/ and /kæptæp/ respectively – a point some time after there is mismatch between the two sets of stimuli. In the following sections the experimental and simulation results for each probe position will be described. In analysing the results of the network simulations, stimulus type corresponds to whether the test sequences contains a short or a long word – equivalent to the prime stimulus factor in the experimental data. Lexical unit refers to whether the activation of a short or long word unit is being measured – analogous to the long or short target word in the priming experiments.

6.4.2. Experiment 2a – AP_1

The main result obtained in this cross-modal priming experiment was that at the offset of a syllable there is sufficient information for listeners to distinguish an embedded word from the onset of a longer competitor. As shown in Figure 6.6a, there was a cross-over interaction between the length of the prime and target such that greatest priming is observed in conditions for which the prime stimuli contains a word of the same length as the target. This pattern is shown by a significant interaction between prime and target length in the difference score analysis with no main effect of either prime or target length.

As can be seen in Figure 6.6b a similar pattern of results is shown for the activation of short and long lexical units in the network. Analysis of variance on the data obtained in the ten trained networks confirms the presence of a significant interaction between stimulus type and lexical unit ($\underline{F}[1,9]=45.39$, $p<.001$) with no main effect of stimulus type

(F<1) and a marginal effect of lexical unit (F[1,9]=4.54, p<.1) suggesting increased activation for short lexical units.

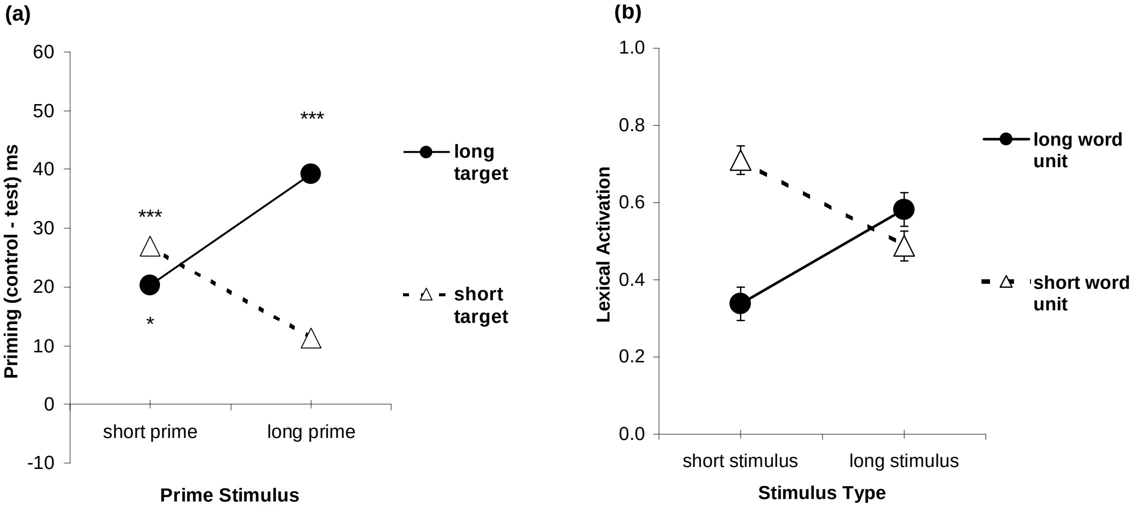

Figure 6.6: Activation of short and long words at *AP*$_1$ – following /kæp/ in *cap* or *captain*. Network stimuli presented with ambiguous duration and prior context to allow disambiguation.

(a) Magnitude and significance of priming in Experiment 2a (* p<.001; * p<.05)**

(b) Mean lexical activation for 10 networks in simulation 4 (error bar = standard error)

Thus the network simulates the main piece of experimental data suggesting that listeners are able to use acoustic cues to word length in the recognition of onset-embedded words and longer competitors. Increased activation is observed for lexical units that match the stimulus. Since the duration cue associated with these syllables is ambiguous in the absence of prior context, in order to display this result the network must be using prior context to disambiguate the input sequences. Thus in the absence of prior context the network predicts that the two sets of stimuli would be indistinguishable at this point. This prediction could be tested in subsequent experiments.

One difference between the experimental results and the simulation is that the network produces a marginally significant increase in activation for short words over long words that is in the reverse direction to the numerical trend observed in the experimental data. This increased activation may reflect the greater number of monosyllabic words in the networks' vocabulary, since the ambiguous duration code is more frequently paired with a short word than with a long word. However, since neither the effect in the experimental data, nor the reverse effect in the model reaches statistical significance, it can still be concluded that there is good overall agreement between the simulation and experimental data.

6.4.3. Experiment 2b – AP_2

At the second alignment point, information in the continuation of the second syllable becomes available in the speech stream. In the cross-modal priming experiments, this increased the amount of priming observed for long targets – especially for prime stimuli that contained the long word. This is reflected in the analysis of priming effects at this probe position, which found main effects of prime and target type (more priming of long targets and more priming from long primes) in addition to the significant interaction between prime and target type. This pattern of results is shown in Figure 6.7a.

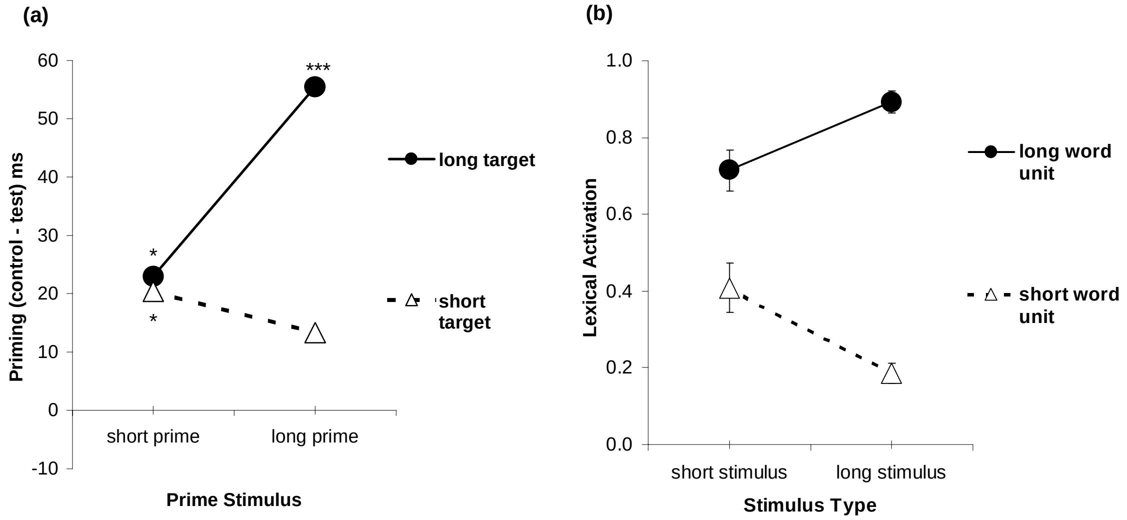

Figure 6.7: Activation of short and long words at *AP*$_2$ – following /kæpt/ in *cap tucked* or *captain*.
(a) Magnitude and significance of priming in Experiment 2b (* p<.001; * p<.05)**
(b) Mean lexical activation for 10 networks in simulation 4 (error bar = standard error)

As can be seen by comparing this graph with Figure 6.7b, the model shows the same interaction between stimulus type and lexical unit as was obtained in the priming data (F[1,9]=25.55, p<.001). However these networks are less successful in simulating the main effects of prime and target type shown in the experimental data, since they show significantly greater activation for long word units, irrespective of prime condition. This effect is confirmed by the significant main effect of lexical unit in the analysis across all ten networks (F[1,9]=51.38, p<.001). Conversely the model does not show greater overall activation for long stimuli than short stimuli (F[1,9]=1.73, p>.1) indicating that while the long word stimuli in the experiment produce greater overall priming (possibly as a result of their reduced ambiguity) the same pattern is not observed in the model.

These discrepancies between simulations and priming data indicate that the model predicts increased activation for long word units for both short word (*cap tap*) and long word (*captain*) stimuli. By comparison, increased priming of long words is only shown where the prime stimulus actually contains a long word. Thus, this comparison of simulations and experiments, suggests that there may be experimental evidence for acoustic cues in word-initial segments that mark word boundaries. Previous experiments (Gow & Gordon, 1995; Nakatani & Dukes, 1977) have suggested that segmental cues in word onsets support the detection of word boundaries. Incorporating these acoustic cues into the model may therefore reduce the effect of garden-path continuations on the activation of long words and improve the networks' simulation of the experimental data. However, without further evidence to support the presence of acoustic cues to word onsets in the stimuli used in Experiment 2, it is premature to alter the input representation of the model.

6.4.4. Experiment 2c – AP_3

The sentences used in Experiments 1 and 2 were designed such that short and long words only diverged in the vowel of the second syllable of the critical stimuli. Consequently, it was expected that the activation of short word hypotheses would increase at the third alignment point, where information in the vowel becomes available to participants. However, the results of Experiment 2c showed that the short word stimuli remained ambiguous at AP_3 as shown in Figure 6.8a. Statistical analysis showed a marginally significant main effect of target type (by participants but not by items) suggesting greater priming of long word targets. There was no main effect of prime type and the interaction between prime and target length was again significant at this probe position.

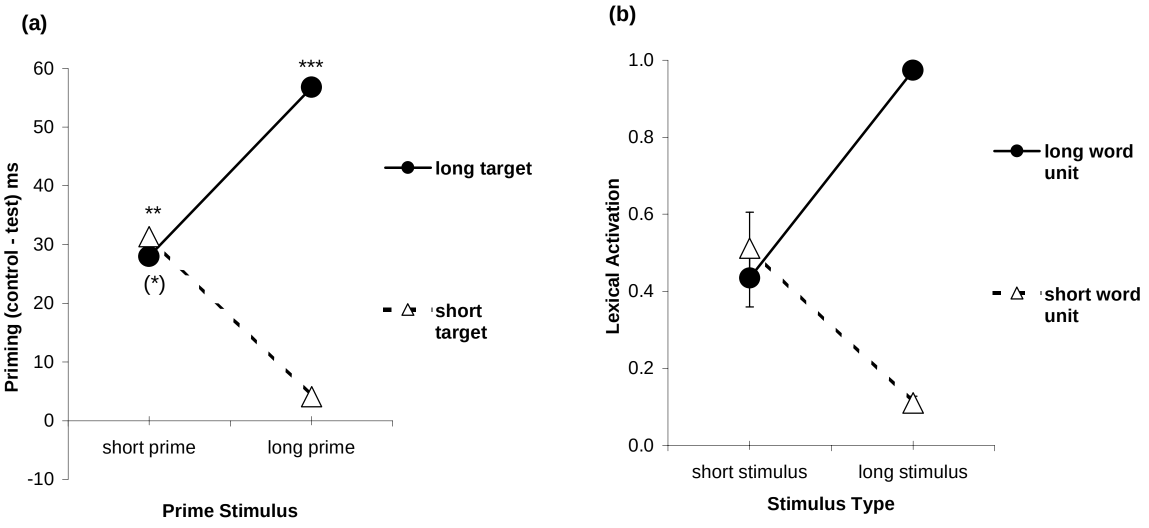

Figure 6.8: Activation of short and long words at *AP*$_3$ – following /kæptuː/ in *cap tucked* or /kæptɪ/ in *captain*

(a) Priming results from Experiment 2c (* p<.001; ** p<.1; (*) p<.1)**

(b) Activation for 10 networks in simulation 4 (error bar = standard error)

As can be seen in Figure 6.8 there is a very strong resemblance between the pattern of priming observed in Experiment 2c and the activation of lexical units in simulation 4. This resemblance is supported by statistical analysis of lexical activations in the network which showed a main effect of lexical unit (F[1,9]=26.83, p<.001) equivalent to the main effect of target type in the experimental data. Effects of stimulus type were non-significant in the network (as for the priming experiment) and the interaction between stimulus type and lexical unit was highly significant (F[1,9]=54.22, p<.001).

Thus, there is good agreement between the network and the priming data regarding the time course of responses to mismatch in the speech stream. Both the recurrent network and the experimental data suggest that despite the presence of mismatch between short stimuli and long lexical items at *AP*$_3$, effects of mismatch are slow to act in reducing the activation of long words.

6.4.5. Experiment 2d – *AP*$_4$

The final probe position tested in the cross-modal priming experiments marked a point at which both short and long word stimuli were expected to be ambiguous. This lack of ambiguity is apparent in the priming effects obtained in Experiment 2d, shown in Figure 6.9a. The only significant effect in the statistical analysis of these data is a cross-over interaction between prime and target length. This reflects the expected pattern – that significant priming is observed only for conditions in which the prime and target match.

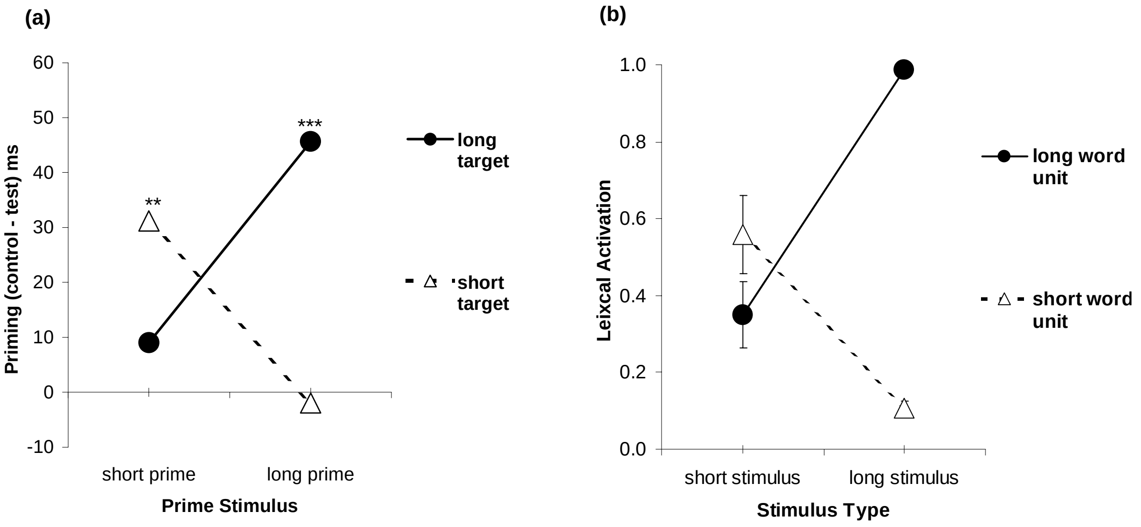

Figure 6.9: Activation of short and long words at *AP*$_4$ – following /kæptuːk/ in *cap tucked* or /kæptɪn/ in *captain*

(a) Priming results from Experiment 2d (* p<.001; ** p<.1)**

(b) Activation for 10 networks in simulation 4 (error bar = standard error)

Once more, visual comparison of the priming data and network activations shown in Figure 6.9 are encouraging. Both graphs suggest that short and long stimuli can be identified at this probe position (though long stimuli appear to be less ambiguous than short word stimuli). However, statistical analysis of network activations do not reflect this apparent similarity. ANOVA shows a significant main effect of lexical unit indicating greater overall activation of long word units (F[1,9]=17.29, p<.01). There is also a marginally significant effect of stimulus type indicating greater activation for long word stimuli (F[1,9]=3.40, p<.1). More reassuringly, the most significant effect in the analysis of the networks' performance is the interaction between stimulus type and lexical unit (F[1,9]=55.11, p<.001). This indicates that despite these discrepant main effects, the simulation does capture the lack of ambiguity of the stimuli at this probe position.

Both of the main effects reported in the model appear to result from the very small amount of variance that is observed in output activations for long word stimuli. As suggested by the invisibility of the error bars on the right hand side of Figure 6.9b all ten networks fully activated the long word units and deactivated the short word units at this position in the long word stimuli[1]. In contrast to this unambiguous activation profile for long word stimuli, short word stimuli are rather more ambiguous at this probe position and hence activations are more variable for these stimuli. However, since priming data are highly variable for both short and long target words, main effects of prime and target type are not observed in the analysis of the experimental data.

6.4.6. Discussion

The final set of simulations reported here demonstrate that simple recurrent networks are able to account for the integration of segmental and supra-segmental cues to the identification of onset-embedded words in connected speech (post-offset mismatch and syllable duration respectively). Despite the limited vocabulary on which the model was trained, statistical analyses of network activations showed many of the same effects that were reported in the cross-modal priming data presented in Chapter 5. Although the exact details of the results at each probe position are sometimes lacking, the network clearly simulates the initial lack of ambiguity of short and long word stimuli. At subsequent probe positions the network also simulates the bias towards long word interpretations followed by the reduction in ambiguity of the short word stimuli when mismatching input is processed.

The network results illustrate the promise of an account, in which lexical activations are proportional to the conditional probability of individual lexical items in the input. However, caveats remain about the representational assumptions that have been incorporated into these networks. In modelling the adaptive processing of input cues analogous to duration, these simulations have focussed on a highly restricted subset of the variables involved in determining the duration of segments and syllables in connected speech. Although the greater complexity of real speech appears to make the task faced by the network more difficult, the addition of more realistic speech input may reduce the absolute degree of ambiguity that is present in these stimuli. If multiple cues to syllable duration were present, it is unlikely that the length of syllables would be as exactly matched as was assumed in the simulations reported here. This discussion illustrates one

[1] The residual activation of short word units represents the likelihood that they will appear as one of the remaining words in the current sequence

limitation of the recurrent network simulations developed here – the highly unrealistic input representation used in the model. Further work is therefore required to investigate whether these networks can incorporate more realistically structured input and output representations.

Representing the speech input

The networks used in these simulations are provided with inputs that represent the duration of segments and syllables independently of the identity of these input segments. This assumption does not hold for real speech since duration, particularly for vowel segments, is strongly affected by the identity of adjacent segments (Klatt, 1976). Furthermore there is evidence from gating studies (Warren & Marslen-Wilson, 1988) that this duration information can contribute to lexical choice. Finally, in many languages, lexical items are contrasted solely by vowel duration.

Since each of these properties contradicts the assumptions used in developing the networks reported here, further work is required to show how differences in duration can be more appropriately represented in the network. However, simulations representing differences in duration as different numbers of duplicated input segments have not been successful. The SRN that was used here fails to process duration coded as duplicated segments in an adapative manner.

An alternative means of processing duration information is therefore required for a complete account of sensitivity to temporal structure in the speech stream. One account of how a network could adapt to differences in the temporal properties of the speech stream is provided by the dynamic rate adaptation networks investigated by Nguyen and Cottrell, (1997). They investigated recurrent networks in which the time delay on their recurrent connections is adjusted to minimise prediction error. These systems can thereby match their processing properties to the rate of the current input. However this process is implemented as an off-line process and thus may not be appropriate in simulating the on-line processing of spoken input.

An alternative account is provided by the entrainment processes implemented in oscillator based networks (Gasser, Eck, & Port, 1999; McAuley, 1994). These systems use discrepancies between the predicted and actual position of metrical beats in the onsets of stressed syllables to alter the rate of oscillation of processing units. Thus systems of these neurons will gradually adapt to the rate of an ongoing sequence.

Either of these approaches provides an account of how a connectionist system can alter its computational properties to compensate for changes in the rate of presentation of the speech stream. Either of these approaches may therefore be required for a more complete model of temporal processing of speech stimuli in connectionist networks.

6.5. General discussion

The networks shown in Simulation 4 are very successful in accounting for the pattern of priming data produced in Experiment 2. However, as is apparent from the small scale of the model, further simulations are required to extend this account to cover more realistically sized vocabularies. This should allow the network to simulate experimental data other than that reported in this thesis. For instance, it may be possible to extend the model to account for a wider range of ambiguous sequences, such as the minimal pairs (e.g. *grey day* and *grade A*) used in the acoustic phonetics literature (see Chapter 2 for further details).

In addition to increasing the model’s coverage of segmental ambiguity, this architecture could also be trained to simulate a wider range of supra-segmental phenomena – such as the difference between metrically stressed and unstressed syllables. In this way it may be possible to simulate results used to motivate the metrical segmentation strategy (Cutler and Norris, 1988). These extensions to the model will further test the probabilistic approach to lexical access and segmentation that has been developed in this thesis.

The recurrent network simulations presented here have shown that a probabilistic approach to the process of spoken word recognition has the potential to simulate many of the results that have previously motivated lexical competition models such as TRACE and Shortlist. These simulations have therefore demonstrated how computational properties of the recognition system such as bottom up activation and inhibition of lexical candidates, competition between lexical items spanning potential word boundaries and adaptive processing of the input can all be learnt by a simple gradient descent algorithm. Further investigations are required to establish if this account is able to simulate a wider range of experimental data.

Even in its current, limited form, the network makes a number of testable predictions for future experiments. The first of these is that the acoustic cues that discriminate syllables of short and long words should be more easily detected with a preceding sentential context than in the absence of such a context. A second prediction is that altering the temporal properties of the critical syllables (or their preceding context) will have an strongly biasing effect on subjects' interpretations of those syllables.

A further prediction made by these simulations is that the segments that follow the offset of an embedded word will affect the ease with which subjects can identify an embedded word. Garden path sequences (such as *"cap tucked"*) in which continuations match a longer lexical item (*captain*) are particularly difficult for the network to identify. In contrast sequences which mismatch with all other lexical items immediately after the offset of the embedded word will be comparatively easy to process. It is this prediction of the network that will be tested in the following chapter.