Lexical Segmentation: the role of sequential statistics in supervised and un-supervised models

Paul Cairns* and **Richard Shillcock*** and **Nick Chater†** and **Joe Levy‡**
*Centre for Cognitive Science, †Department of Psychology,
‡Human Communication Research Centre,
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, U.K.
Tel: +44 31 650 4667
FAX: +44 31 650 4587
pcairns@cogsci.ed.ac.uk[1]

Abstract

The use of transitional probabilities between phonetic segments as a cue for segmenting words from English speech is investigated. We develop a series of class-based n-gram and feature-based neural network models that enable us to quantify the contribution of low-level statistics to word boundary prediction. Training data for our models is representative of genuine conversational speech: a phonological transcription of the London-Lund corpus. These simple models can be purely bottom-up and hence valid bootstrapping models of infant development. We go on to demonstrate how the boostrapping models mimic the Metrical Segmentation Strategy of Cutler and Norris (1988), and we discuss the implications of this result.

Introduction: The Segmentation Problem

The majority of word boundaries in normal speech are not reliably marked by acoustic breaks. Segmentation is the process of dividing up the continuous input speech stream into linguistically and psychologically significant units that can be used to access meaning. Segmentation and recognition appear to stand in a chicken-and-egg relationship: the extraction of a meaningful unit presupposes recognizing what that unit is; but recognition only seems to be possible once segmentation has been carried out.

Broadly speaking, there are two ways in which to break out of the segmentation ⇔ recognition circle for the adult listener. The first is to use an interactive approach, in which the system puts forward tentative hypotheses concerning segmentation and recognition on the basis of lexical information, and reinforces hypotheses that fit together (e.g. Marslen-Wilson and Welsh (1978); McClelland and Elman (1986)). The second approach is to attempt to find reliable cues for segmentation, which are independent of the identity of what is being segmented. According to this second approach, segmentation can be carried out bottom-up, and its output fed on to later recognition processes.[2]

In acquisition, an isomorphic segmentation problem exists, except that the goal of segmentation is not lexical access but lexical compilation. Here it seems more likely that bottom-up cues are used at least initially, since in the earliest stages of acquisition the infant has no lexicon with which to segment interactively. Although it has been suggested (e.g. Suomi (1993)) that words spoken in isolation could be stored and thereafter used to aid segmentation interactively, there is no quantitative evidence that this is feasible.

The bottom-up cues that may be used to aid segmentation can be divided into three main groupings: (i) Acoustic/phonetic juncture markers or pauses (ii) Prosodic marking that specifies the initial portion of a word, given a presyllabified input. (iii) Distributional cues: for example differing probabilities of certain phonological sequences at various points in the speech stream (*phonotactics*). The first type of cue has been studied in the phonological and speech recognition literature (Lehiste (1971)). The second approach has been thoroughly investigated by Cutler and colleagues (Cutler and Norris (1988); Cutler (1993); Cutler and Butterfield (1992)). Her *Metrical Segmentation Strategy* (henceforth *MSS*) holds that when a strong vowel is heard, a boundary is hypothesized at the beginning of the syllable of which the vowel is nucleus. Although originally a theory of adult behaviour, there has been recent discussion of how the MSS could be acquired (see Cutler et al. (1992); Otake et al. (1993)), and also work that seems to demonstrate a sensitivity to metrical patterns in infants of 9 months (Jusczyk, Cutler and Redanz (1993)). The third type of cue we refer to as *phonotactics*, by which we mean constraints on the segmental phonological structure of words and syllables. It is generally the case that sequences of segments are more constrained within words than across word boundaries. Thus, the sequence /ŋð/ is only licenced in English if there is a morpheme boundary between /ŋ/ and /ð/. However, phonotactics do not have to be absolute constraints, probabilistic structure is present too: thus the sequence /z ə/ is very common across word boundaries, but much less common word-internally. The role of phonotactics has been studied in the speech recognition literature (see Harrington, Watson and Cooper (1988)), but has received little attention in the domain of psycholinguistics.

In this paper, we present two bottom-up statistical models which can be applied to adult behaviour, and infant development, respectively. Our models exclusively use phonotactic information. This is not because we believe that phonotactics is the only information source that listeners use in segmenting speech. Rather, we hope to quantify precisely the possible contribution of phonotactic information to a more complete model of segmentation which would integrate information from various sources.

[1]This work was supported by the U.K. Economic and Social Research Council (ESRC). Grant number: R000 23 3649

[2]For a model detailed discussion of the topic see Cairns et al. (1994).

A Phonological Re-transcription of the London-Lund Corpus

In order to be admissible as support for the bottom-up approach, the data from which a model of segmentation is derived must be representative of real speech. Accordingly, we present a large corpus of phonologically transcribed speech.[3] The London-Lund corpus (*LLC*) is a body of English conversation transcribed orthographically and available on-line. Because of its size (around 460,000 words) an automatic method was developed for its phonetic transcription. First, the words are replaced by their phonemic citation forms using an on-line dictionary. Then, these forms are input to a set of re-write rules that introduce phonological alternations into the string (e.g. assimilation, vowel reduction). None of the rules uses word boundary information to specify its context of application. The output from the rule-set is a corpus of 1.5 million phonetic segments.

It is, of course, impossible to recreate the original speech data, but this method has two advantages: First, we need a large corpus of conversational speech if its statistics are to be representative — at present there is no comparably large corpus with a genuine phonological transcription; Second, this method provides a higher-order approximation to genuine data, when compared with a corpus derived from a phonemic dictionary in combination with word frequency counts. Thus, our data is representative of the distribution of strings of closed-class words such as `if I can`. Any adequate model of segmentation must cope with such input. Two important characteristics of our corpus are: (1) All rules for co-articulation apply equally inter- and intra-lexically. (2) The data is very noisy with frequent repetition, hesitation, errors, etc. Because of these facts, we consider the data to represent a "worst case" for testing models of segmentation, in that if segmentation is possible with this data, then the inclusion of pauses, prosody, and some phonetic/acoustic cues can only serve to improve performance.

In the experiments reported here, we will use two versions of the corpus. Corpus A has word boundary markers, while Corpus B has **no** explicit word boundary marking.

N-gram Models of Adult Segmentation

Using corpus A (corpus **with** word boundary markers), the prior for all bigrams: $\{\langle p^1, p^2\rangle \mid p^1, p^2 \in P\}$ was calculated, where the pair were either word internal, or straddled a word boundary (P is the set of all phonemes). We use the ratio of the prior that a pair $\langle p^1, p^2\rangle$ occurs across a boundary to the prior that it occurs within a word, denoted $p_{across}(\langle p^1, p^2\rangle)/p_{within}(\langle p^1, p^2\rangle)$ to decide when to propose a boundary. When this ratio rises above a certain cutoff point we insert a boundary. When the cutoff is set high, the performance of this model tends toward the behaviour of the deterministic n-gram model of Harrington, Watson and Cooper (1988). Note that because we are specifying word boundary location in the n-grams, the model is *supervised*, and therefore is not applicable as a model of development. However, once it is trained, the model is strictly bottom-up in operation.

The results of running this segmentation algorithm on a 10,000 phoneme (approx. 2,800 word) test stretch of the same corpus can be seen in Figure 1 where we plot the probability of a hit versus a false-alarm as the cutoff is varied.[4] Selecting different cutoff points can provide performance such as detection of 45% of the boundaries in the test stretch with a hits:false-alarms ratio of 45:1, or 66% of all boundaries with a hits:false-alarms ratio of 9:1. Where exactly to place the cutoff point is a question that depends on our theory of how much of a problem false-alarms and misses pose for the human processor, which will reflect assumptions about processor modularity, parallelism in processing, and so forth. However, one can measure, in a pre-theoretical manner, how well the segmentation algorithm performs by taking an information theoretic measure such as *mutual information* at each cutoff point, and choosing the cutoff at which this measure is maximized. In effect, the mutual information measure tests whether the general shape of the distributions of boundary points is the same for the segmentation algorithm and the true stretch of segmented corpus, as well as the extent to which the individual decisions match. At the mutual information maximum the detection rate is 75% with a hits:false-alarms ratio of 4.7:1.

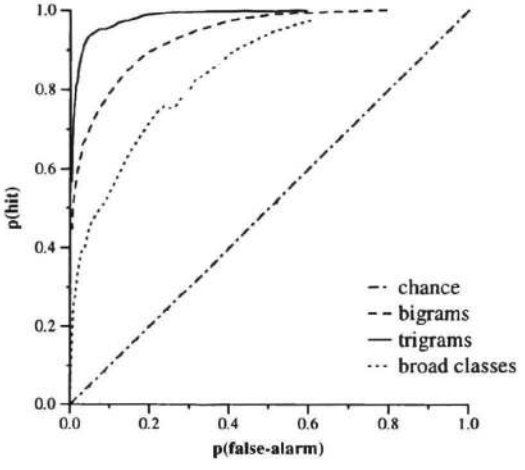


Figure 1: *ROC* (Receiver Operating Characteristic) graph for: n-gram segmentation performance (corpus with boundaries marked). The different curves show the results for bigrams, trigrams, and a trigram model where the transcription was in terms of the 6 broad phonetic classes.

Further improvement can be made on this performance by using trigrams rather than bigrams. We collected the priors of all triples: $\{\langle p^1, p^2, p^3\rangle \mid p^1, p^2, p^3 \in P\}$ that were either word-internal, or had a boundary between p^1 and p^2, or had a boundary between p^2 and p^3. However, now we have two ratios p_{across}/p_{within} where p_{across} can correspond to the

[3] A more detailed description of the corpus can be found in Shillcock, Cairns, Levy, Chater, and Lindsey (1993)

The full-scale approach concomitant with use of a corpus can also bring benefits in other psycholinguistic domains: see Chater, Shillcock, Cairns, and Levy (1993).

[4] A *false-alarm* occurs when the model posits a boundary when no boundary is actually present. A *hit* is when the model successfully detects a real boundary.

sequence $\langle p^1, \#, p^2, p^3 \rangle$ or $\langle p^1, p^2, \#, p^3 \rangle$. As a first step, we simply took the mean of the two ratios, and moved the cutoff point relative to this figure. There are further complications that arise through the use of trigrams: the tendency to over-segment when there are one and two-letter words in the input. A remedy for this problem is to have a list of permissible one-phoneme words (for present purposes just /ə/ and /ou/), and not to license segmentations that create one-phoneme words not on this list. Having done this, the results for the segmentation of the same test stretch of corpus as before are shown in Figure 1. The trigrams show a considerable improvement on the bigram figures, with performance ranging from detection of 57% of the boundaries with a false-alarm rate of 65:1, to the mutual information peak at 93% detection with 9:1.

The algorithm does indeed show some over-segmentation of inflectional forms as Harrington et al. realized would happen. However, as can be seen from the results these cases are really quite rare in normal conversational speech. Another common error is to over-segment words which begin with a weak vowel, thus /tɔk#ə#baut/ for `talk about`, though once again such cases are rare. In fact, this latter type of error — where a word boundary is spuriously inserted before a strong vowel — is very common in human slips of the ear (see Cutler and Butterfield (1992)), so one could interpret this as being a feature of the phonotactic model.

These figures would seem to indicate that the problem of segmentation is not really such a problem after all. However, our result must be qualified by noting its possible reliance on a detailed and unambiguous phonemic input string, something which in all probability is not obtainable either in an Automatic Speech Recognition (henceforth *ASR*) system, or in human listening. In real speech, phonemes are realised with numerous variations in both time and quality. Of course, the solution that many phonologists and psycholinguists take is to assume that such issues can be resolved at a lower level, and will not impinge on higher level processes, however there seems to be little evidence that this is the case.

We tested the reliance of these results on a clear transcription using a full phonemic inventory, by following Zue and colleagues (e.g. Huttenlocher and Zue (1983)) who have used transcriptions of speech where each segment is placed in one of six broad phonetic classes which are more reliably identified by an ASR system. We re-transcribed our corpus in terms of the six classes (*Stop, Nasal, Weak Fricative, Strong Fricative, Glide/Liquid,* and *Vowel*) and once again constructed a trigram model using exactly the same procedure as before. Not surprisingly, the performance was degraded when compared to the fully transcribed trigram model (see Figure 1). However, absolute performance is still good, with a mutual information peak at 74% detection with 1.5:1 hits:false-alarms.

To summarise, the supervised trigram technique provides a powerful model of adult behaviour, but its effectiveness is proportional to the detail of the input phonological transcription.

Connectionist Modelling of Segmentation Acquisition

In addition to the supervised n-gram models of adult segmentation, we have also constructed un-supervised n-gram models in which no word boundary marking is employed in construction (i.e. # is *not* a member of the symbol set) using corpus B. These models, unlike the supervised n-gram models just described, are admissible as models of infant development because no top-down information is employed during training. However, such models inherently employ phonemic categories, and hence cannot be used to address the pre-categorial phase of infant development. Therefore we developed a neural network model that is feature-, rather than category-based. We consider a feature-based representation to be one step closer to the genuine speech signal.

Network Training

The network has a recurrent, self-supervised, architecture (see figure 2). The task is to echo the current slice of input, remember the previous, and, most importantly for this paper, to predict the next. As input to the model, we translated corpus B described above into a nine-bit binary feature vector representation where the features are taken from the Government Phonology scheme of cognitive elements (see Harris and Lindsey ((in press)); Kaye, Lowenstramm and Vergnaud (1985); Shillcock et al. (1992)).[5] Noise is added to the input by flipping features from 0 to 1 (or vice versa) with a certain probability, in order to encourage the network to rely on sequential information (i.e. if the current segment is obscured, then the net will have an incentive to use the local phonetic context to recover its identity). The net is trained using Back-propagation Through Time (*BPTT* — see Rumelhart, Hinton and Williams (1986)), a steepest descent procedure, and a cross-entropy error measure (see Hinton (1989) — cross entropy is a good measure to use when one wishes to interpret continuous valued outputs as probabilities of binary decisions). Training comprises two passes through a training stretch of the corpus one million phonemes in length (with different noise on each pass), thus two million phonemes in total. The learning rate is decayed as training progresses.

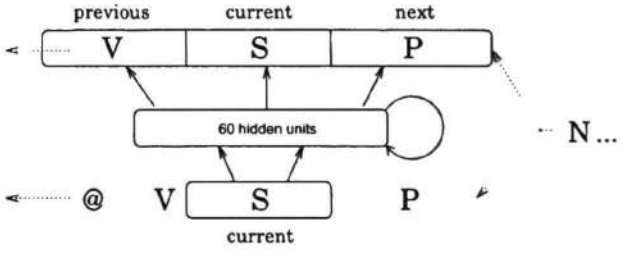


Figure 2: The Network – The solid arrows between layers indicate complete connectivity with modifiable uni-directional links. The dotted arrows show how the input corpus arrives over time to specify the input and output target.

Network Segmentation

The rationale used in postulating boundary points follows from what we know about phonotactics. From the network's point of view, lack of constraint in phonotactic structure (high

[5] Work by Williams and Brockhaus (1992) has shown how the government phonology elements can be automatically extracted from the speech stream, so we have reason to believe that coding in this way represents a step further towards ecological validity.

entropy in information theoretic terms) will make the next segment difficult to predict. If prediction is hard, then error will be high. Thus, boundaries are proposed at peaks in the error score on the prediction output units (marked `next` in Figure 2).

The model was tested by providing as input a noise-free 10,000 phoneme (about 2,700 words) stretch of corpus, and measuring the Cross Entropy error on the prediction subgroup of the output units. This yields a variable error signal in which we define a "peak" by placing a cutoff point at varying numbers of standard deviations above the mean. The effects of choosing increasingly more stringent cutoff points can be seen in Figure 3 where we plot how the hit and false-alarm rates vary with the cutoff point. At the cutoff that maximizes the mutual information, 21% of the boundaries are correctly identified with a hits:false-alarms ratio of 1.5:1.

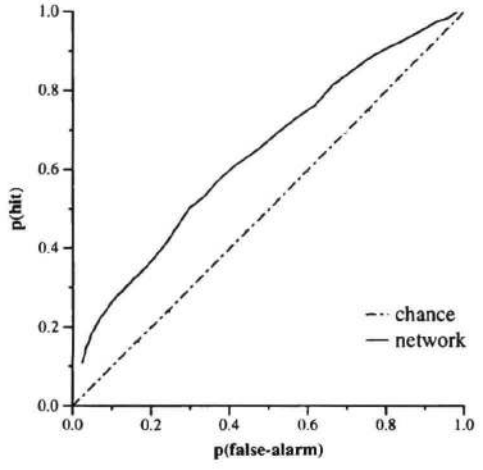


Figure 3: *ROC* graph of network segmentation performance (corpus with **no** word boundary marking).

In the following sections we evaluate the significance of these results by comparison with a random segmentation algorithm which was averaged over five different runs. This algorithm was designed to yield a distribution of "word" length similar to that of the network. We consider this to be a more stringent test of the network's performance than comparison with a random segmentation algorithm that uses a uniform distribution. The difference in performance is highly significant: $\chi^2_{(1)} = 216.8, p < 0.001$.

Although network performance peaks with correct identification of about one in five boundaries in the test corpus, there is a sizable proportion of false alarms at this cutoff (i.e. cases in which the network predicts a boundary when in fact there is none). It may well be that although the false-alarms do not actually correspond to existing boundaries in the test stretch, they are actually plausible guesses based on the low-level data that is the only information source available to the model. We tested this hypothesis by examining the phonotactic acceptability of the boundaries that the model postulates, defined by the legality of the sequence of segments over the postulated boundary. Thus the sequence /tp#ra/ is a phonotactically malformed boundary postulate, while /pt#ra/ is well-formed. We found that false-alarm boundaries of the network are much more likely to be phonotactically well-formed than those of the random case (for the initial boundaries: $\chi^2_{(1)} = 221.8, p < 0.001$, while for the final boundaries $\chi^2_{(1)} = 119.1, p < 0.001$).

In summary, phonotactics provide a fairly weak source of information for the bootstrapping of segmentation, but the cumulative effect of such information may well be useful in the initial phases of compiling a lexicon.

Network performance and the *MSS*

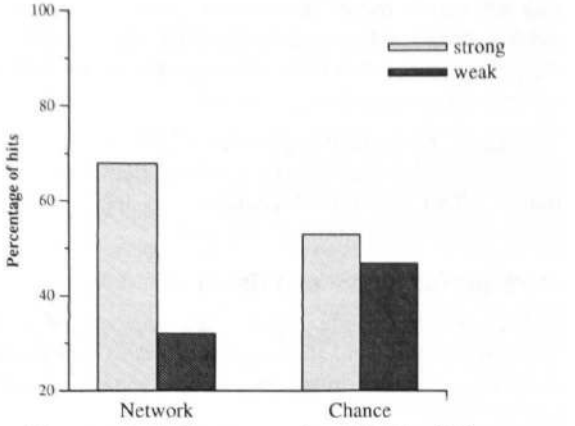
In this section we provide a qualitative analysis of network segmentation and present the surprising result that there is a statistical basis for the emergence of the *MSS* in our purely bottom-up model.

We investigated the performance of the model by counting the instances in which a boundary is correctly postulated before a strong or weak syllable. The definition of weak and strong is not trivial however. While the status of schwa (/ə/) as a weak vowel is inherent, other short vowels such as /a/ and /ɪ/ can be either metrically strong or weak depending on context (this version of the corpus is not transcribed with metrical markings). As an operational definition of *strong* and *weak* we took the lax vowels /ə/,/ɪ/, and /ʌ/ to be weak, and all other monothongs and dipthongs to be strong. Because some of the instances of /ɪ/ and /ʌ/ which we classify as weak will actually be strong, if anything this will tend to artificially boost the number of weak classifications. Given this criterion, in the 2,700 word test set 53% of the words are strong-initial.[6] The network performance is proportionally skewed towards successful detection of strong-initial words to a striking degree (see Figure 4a, $\chi^2_{(1)} = 77.2, p < 0.001$). A similar result was obtained when we changed the definition of *weak* to just /ə/ ($\chi^2_{(1)} = 70.4, p < 0.001$). A natural conclusion to draw is that the model is segmenting more before open-class words, and examination of the totals of hits before open- as opposed to closed-class shows that this is the case. The initial portions of open-class words are much more likely to be detected than beginnings of closed-class items ((see Figure 4b, $\chi^2_{(1)} = 14.0, p < 0.001$). Note also that the boundaries with which the model has most difficulty are the closed-closed boundaries, thus strings of closed-class words such as `up to the` are less likely to be segmented than strings of open-class items.

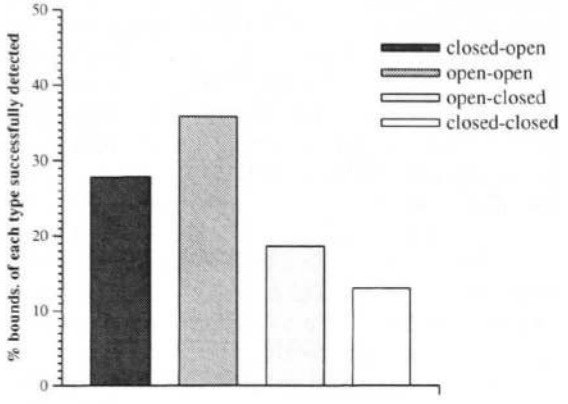
When we consider the contiguous pairings of these individual segmentations — the words that emerge from the network — the same pattern is evident. A word count of the LLC revealed that 65% of all items were closed-class, so one would expect that this ratio would hold in network output, all other factors being equal. While the network does not segment more whole words from the test stretch than it would by chance (showing that the model does not develop a lexicon), of the correctly extracted tokens 41% are closed-class. This is significantly different from the random segmentation performance: $\chi^2_{(1)} = 19.46, p < 0.001$.

So, our network produces segmentations which broadly mimic the pattern predicted by the MSS, yet the net is not retrodictive in the way that the MSS is: Crucially, the nuclear

[6] This is representative of the proportion for real speech, see Cairns et al. (1994).



a) Segmentation before strong and weak initial syllables.



b) Segmentation across different boundary types.

Figure 4: Network segmentation performance mimics the *MSS*.

vowel of the initial syllable is not visible to the network when it makes a segmentation decision (recall that boundaries are inserted on the basis of the ease of prediction of the first segment in the word — generally a consonant). Thus, this model does not need to posit that the Strong/Weak distinction has *a priori* perceptual salience for the infant. The reason why our network exhibits this pattern of results is simply that the initial segments of strong-initial, open-class words tend to be less predictable than those of closed-class words. However, we would emphasize that phonotactic information is a relatively weak predictor, and that it is unlikely that a purely phonotactic approach could enable the infant to acquire a lexicon. Rather, we see phonotactic information as being a possible method for bootstrapping of the MSS, which is a more robust and reliable tool for lexical acquisition. This bootstrapping could be mediated by sensitivity to the correlation of the boundaries that phonotactics predict with metrical structure.

Adding categorial knowledge

The results from the previous section were obtained by segmenting with raw scores that were not normalized for phoneme type. This can be seen as simulating the phase of infant development in which phonemic categories, and information about their frequencies, are not available to the infant. However, we know that towards the end of the first year of life the child's phonological space is becoming structured with phonemic categories (e.g. Kuhl (1983); Werker (1993)). Therefore, we decided to mimic the effect of this phonemic restructuring in our model, to see if the qualitative pattern of segmentations would remain constant.

We carried out the same segmentation procedure as before, but this time normalizing the network error scores for phoneme type. We found an entirely different pattern of results with respect to strong-syllables and word class than before: in general the network no longer mimicked the *MSS*. Segmentation before strong as opposed to weak syllables was not significantly different from chance: $\chi^2_{(1)} = 0.387, p > 0.05$. Neither was segmentation before open as opposed to closed-class items: $\chi^2_{(1)} = 0.035, p > 0.05$. Furthermore, using phoneme-normalized scores, 78% of correctly extracted word tokens were closed-class, in contrast to the 41% with raw scores. This figure once again differs significantly from the expected distribution: $\chi^2_{(1)} = 8.07, p < 0.005$, except that now it is the closed-class items that are favoured, rather than the open-class.

The intuitive explanation of why segmentation behaviour should change in this way when scores are normalized is that closed-class words, because they are most frequent in the language, also contain the most frequent phonemes. Therefore, the network will predict these phonemes more easily than ones which do not occur often in closed-class words. Because predicting these segments is easier, errors are lower. Hence normalizing for phoneme type will augment the error scores for phonemes that most often occur in closed-class words, and effectively increase the probability of boundaries being proposed before such segments.

Summary and Discussion

First, as regards adult modelling, we have shown that probabilistic n-gram models can be extremely powerful word-boundary detectors, but are reliant on the quality of the phonemic transcription. Therefore, phonotactic information may have a critical role to play in adult segmentation. With reference to how such phonotactic information could come to be encoded in the language processor, we would favour a model in which correlations between successfully activated lexical items, and sequences of segments in a phonological buffer would serve to strengthen or weaken particular n-grams.

Second, we have provided a computational underpinning to the claim that low-level phonotactics could be used by a neonate as a cue for initially breaking up the continuous stream of input speech. Note that experimental evidence shows that infants **are** sensitive to the sequential statistics of natural language (see Jusczyk et al. (1993)).

Moreover, we have given an account of how the *MSS* could arise without recourse to positing metrical information as part of a genetic endowment. The network segmentation performance was significantly biased in favour of detecting open-class words that have strong initial syllables.

Furthermore, we have shown that our model's mirroring of the *MSS* disappears when we add knowledge about the frequencies of individual phoneme categories — detection of closed-class words becomes favoured.

We see the overall picture of the role of phonotactics that

emerges from these results as follows: Initially, phonotactics could provide initial segmentations from which the *MSS* could be induced in the pre-categorial infant. Once the *MSS* is in place, and the infant's phonological space comes to be structured with the phonemic categories of English, then the *MSS* would pick out the open-class words, while phonotactics could help in isolating the closed-class items. This raises the possibility of a critical period for realization of the *MSS*: If we assume that categorial knowledge is pervasive after a certain stage of development, then the utility of the phonotactic strategy for bootstrapping the MSS is only visible in the pre-categorial phase.

References

Cairns, P., R. Shillcock, N. Chater and J. Levy (1994) Bootstrapping word boundaries: a bottom-up corpus-based approach to speech segmentation. Submitted to *Cognitive Psychology*, currently being revised.

Chater, N., R. Shillcock, P. Cairns and J. Levy (1994) Bottom-up explanation of phoneme restoration. November 1993, Centre for Cognitive Science, University of Edinburgh. Submitted to Journal of Memory and Language, currently being revised.

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

Cutler, A. (1993) Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research* **22**(2), 109–131.

Cutler, A. and S. Butterfield (1992) Rhythmic cues to speech segmentation - evidence from juncture misperception. *Journal of Memory and Language* **31**(2), 218–236.

Cutler, A., J. Mehler, D. Norris and J. Segui (1992) The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology* **24**, 381–410.

Harrington, J., G. Watson and M. Cooper (1988) Word boundary identification from phoneme sequence constraints in automatic continuous speech recognition. In *Proceedings of the 12th International Conference on Computational Linguistics*, pp. 225–230.

Harris, J. and G. Lindsey ((in press)) The elements of phonological representation. In J. Durand and F. Katamba, eds., *New Frontiers in Phonology*. Harlow: Longman.

Hinton, G. E. (1989) Connectionist learning procedures. *Artificial Intelligence* **40**, 185–234.

Huttenlocher, D. P. and V. W. Zue (1983) Phonotactic and lexical constaints in speech recognition. *Proceedings of the Third National Conference on Artificial Intelligence* 172–176.

Jusczyk, P. W., A. Cutler and N. J. Redanz (1993) Infants' preference for the predominant stress patterns of English words. *Child Development* **64**, 657–687.

Jusczyk, P. W., A. D. Friederici, J. M. I. Wessels, V. Y. Svenkerud and A. M. Jusczyk (1993) Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* **32**, 402–420.

Kaye, J. D., J. Lowenstramm and J. R. Vergnaud (1985) The internal structure of phonological elements: A theory of charm and government. *Phonology Yearbook* **2**, 305–328.

Kuhl, P. K. (1983) Perception of auditory equivalence classes for speech in early infancy. *Infant Behaviour and Development* **6**, 263–285.

Lehiste, I. (1971) The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America* **51**(6 (2)), 2018–2024.

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

McClelland, J. L. and J. L. Elman (1986) Interactive processes in speech perception: The TRACE model. In J. L. McClelland and D. E. Rumelhart, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2: *Psychological and Biological Models*, pp. 58–121. Cambridge, Mass.: MIT Press.

Otake, T., G. Hatano, A. Cutler and J. Mehler (1993) Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language* **32**, 258–278.

Rumelhart, D. E., G. E. Hinton and R. J. Williams (1986) Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1: *Foundations*, pp. 318–362. Cambridge, Mass.: MIT Press.

Shillcock, R., P. Cairns, J. Levy, N. Chater and G. Lindsey A statistical analysis of an idealized phonological transcription of the London-Lund corpus.

Shillcock, R., G. Lindsey, J. Levy and N. Chater (1992) A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pp. 408–413. Distributed by Lawrence Erlbaum Associates, Hillsdale, N.J.

Suomi, K. (1993) An outline of a developmental model of adult phonological organization and behaviour. *Journal of Phonetics* **21**, 29–60.

Werker, J. F. (1993) Developmental changes in cross-language speech perception: Implications for cognitive models of speech processing. In G. T. M. Altmann and R. C. Shillcock, eds., *Cognitive Models of Speech Processing: The Sperlonga Meeting II*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Williams, G. and W. Brockhaus (1992) Automatic speech recognition: a principle-based approach. In A. Göksel and E. Parker, eds., *Working papers in linguistics and phonetics* 2, pp. 371–401. London: School of Oriental and African Studies.