Models of Language Acquisition

Inductive and Deductive Approaches

Edited by

PETER BROEDER
AND JAAP MURRE

OXFORD
UNIVERSITY PRESS

6

Statistical and Connectionist Modelling of the Development of Speech Segmentation

RICHARD SHILLCOCK, PAUL CAIRNS,
NICK CHATER, and JOE LEVY

6.1. Introduction

The speech signal is typically continuous; only a minority of word boundaries are marked by any recognizable acoustic cue such as a pause. The continuous nature of speech poses a problem for the adult speaker of the language, in that processing the signal requires a complete parse into words yet any string of more than a few segments is locally multiply ambiguous: given only a phonetic transcription, most words contain other words, in the way that *curtain* contains *cur*, or *floor* contains *or*. Segmentation strategies are available to the adult listener that are not given to the infant, as the former possesses both a lexicon containing the phonological specification of the words of the language, and a knowledge of the syntax and semantics of the language. For instance, the adult listener may recognize a word before its acoustic offset and hence may be able to predict the end of the current word and the start of the next; indeed, this strategy featured explicitly in one early model of word recognition (Cole and Jakimik 1980). The adult listener may be able to recruit syntactic knowledge to predict and identify closed-class words (the short grammatical, or function words) (Shillcock and Bard 1993), and hence identify their boundaries too. The infant, faced with the speech sounds of an unknown language, is unable to draw on such knowledge, yet over the first two years of life individual words are isolated in comprehension, stored and begin to be deployed in production. This chapter is concerned with the nature of the information that the infant might exploit to obtain a foothold on the segmentation problem.

We investigate the nature of the speech segmentation problem faced by the infant and we assess a range of 'statistical' solutions based on the principle that all that the infant brings to the problem is a general-purpose capacity to induce the very local statistical structure of sensory input. The simplest form of this statistical strategy asserts only that the sequence of segments *within* words and syllables is more constrained than the sequence of segments *between* words, as evidenced by the sonority hierarchy for instance, so that an unusual transition between two segments is likely to correspond to a word boundary.[1] This is clearly not a sufficient strategy: it is likely to miss some word boundaries and to divide some complex words into syllables. Nevertheless, as we will show below, such a strategy can potentially reveal the usefulness of other, more complex strategies and can contribute to a conspiracy of soft constraints that effectively solve the segmentation problem. Because this simplest statistical strategy assumes very little on the part of the processor, it may play a more important role during language acquisition than it plays in the final, adult repertoire of types of information relevant to segmentation. In contrast to the range of category-types and representational levels required by other types of segmentation information used by the adult listener, the simplest statistical strategy requires only a small number of phonological primitives, together with a general sensitivity to statistical structure—something which seems to be the brain's forte.

In the speech segmentation literature, such statistical models of segmentation are judged to be of very limited value, in spite of the desirability of the minimal assumptions of such models. Cutler *et al.* (1992) refer to studies by Harrington, Watson, and Cooper (1989) and Briscoe (1989) as grounds for not relying on simple models based on phonotactic constraints: such strategies require a very reliable phonetic transcription of the speech stream and are not robust against degradation in the quality of this information. Further, even when given an accurate phonological transcription, purely phonotactic models have been reported as not performing particularly successfully. For instance, Harrington, Watson and Cooper (1988) found that using a dictionary to assess the distribution of different phoneme trigrams within single words, as opposed to straddling a word boundary, gave 37 per cent of the boundaries in their test set of sentences, with one erroneous boundary for every eight correct boundaries identified (a 'hits:false-alarms ratio' of 8:1). We will demonstrate that this pessimism concerning low-level statistical models is not entirely warranted, and that a statistical approach to segmentation may play a significant role both in language acquisition and in adult speech processing.

6.2. Corpus-based Research

Our research has involved the intensive study of a large corpus of transcribed conversation. Over the last few years the availability of speech and text corpora together with the advance in computing power has made the corpus-based analysis of language processing increasingly feasible and attractive. A naturalistic corpus represents, at some level of description, a 'full-scale' or comprehensive approach to the phenomenon: it forces us to deal with the full repertoire of phonological segments, for instance, and it confronts us—in the case of English—with the complete range of closed-class words in their naturally occurring distribution both with open-class words (nouns, verbs, adjectives, adverbs) and with other closed-class words. The larger the corpus, the more it tends to a realistic representation of the frequency of its linguistic constituents, at all levels. A purely dictionary-based approach to modelling speech processing cannot reveal the distribution of short strings of closed-class words, such as *in the*, *of a*, or *out of the* that are so pervasive in conversational English.

The work we describe employs the London Lund Corpus (LLC) (Svartvik and Quirk 1980), the largest available corpus of English conversation and one which permits large-scale statistical investigation. The corpus is replete with repetitions and false starts of normal conversation, and presents us with a formidable segmentation problem.

6.3. Representing the Speech Signal

The LLC is publicly (electronically) available in orthographic form, with some prosodic marking in addition. The ideal corpus for the work we describe would be a corpus of speech with a genuine phonological transcription. Since this is currently not feasible given the size of corpus required (the version of the LLC we used contains some 460,000 words), the closest approximation is to generate an idealized phonological transcription from the LLC's orthographic transcription, our claim being that the inaccuracies inherent in this approach are more than outweighed by the insights that the general approach allows. We have described in more detail elsewhere the complete process of retranscription (Shillcock *et al.* 1993; Cairns *et al.* 1997). First each orthographic word in the corpus (minus the prosodic marking, punctuation and annotations) was automatically replaced by its citation-form phonological transcription, as given in an electronically available phonological dictionary. The closed-class words were given transcriptions suitably phonologically reduced in appropriate contexts. The spaces between words (including marked pauses) were then eliminated, and a limited number of rules applied to the resulting continuous stream to stimulate the effects of coarticualtion—specifically assimilation and non-release of certain stops—between adjacent segments (N.B. these rules were applied with no knowledge of word boundaries in the stream). As an example of a short stretch of the final continuous stream of segments, corresponding to the words *I won't be* the corpus contained /ə w ou m p˺ b i/, in which the /p˺/ is an unreleased /p/.

The resulting idealized phonetic transcription is a rather abstract representation of any original speech signal. Phonetic symbolic representations are

[1] Other researchers have developed information-theoretic approaches to segmentation concentrating on finding frequent sequences and parsing the input in terms of these sequences (see for example Brent 1993; Redlich 1993).

convenient but, at least in formal terms, are inadequate in many ways (see for example Harris and Lindsey 1993). For our modelling of segmentation information, many underlying relationships will be lost by a system which employs more than 40 different categories. Accordingly, we converted the idealized phonetic transcription into a feature-based one (Shillcock *et al.* 1992) grounded in current advances in Government Phonology (Kaye, Lowenstamm, and Vergnaud 1985; Harris and Lindsey 1993). This phonological theory employs nine phonological primitives, or 'elements', defined as follows:

A: oral cavity openness; alone, the vowel quality of *palm*
I: palatality; alone, the vowel quality of *see*
U: labiality; alone, the vowel quality of *boot*
?: occlusion; abruptness; alone, glottal stop
h: aperiodic energy; alone [h]
N: nasality
R: apicality/coronality/coronal formant locus
@: velarity/centrality
H: voicelessness.

TABLE 6.1. Example transcriptions of three segments into Government Phonology elements

Segment	Elements								
	?	h	U	N	R	@	H	I	A
p (pat)	1	1	1	0	0	0	1	0	0
t (tap)	1	1	0	0	1	0	1	0	0
k (cat)	1	1	0	0	0	1	1	0	0

Table 6.1 illustrates the way in which each phoneme may be represented as an aggregate of features, occupying a single timeslice. For the full inventory, see Shillcock *et al.* (1992). Affricates, diphthongs, and long monophthongs are decomposed into two consecutive segments, as illustrated in Table 6.2.

Within the theory of Government Phonology, the elements are seen as being closer to perceptual considerations as opposed to production ones, in contrast to some other phonological theories; indeed some of the elements are ascribed characteristic speech spectogram instantiations. The elements are seen as universal phonological primitives. Thus, Government Phonology elements suit our

TABLE 6.2. Example expansion of affricates and long monophthongs

Segment	Expansion
tʃ (chair)	tʃ
dʒ (germ)	dʒ
əː (bird)	əə

current goals well, in representing a principled, coherent, low-level idealization of the speech stream. Note, finally, that our transcription of the corpus remains particularly abstract in the temporal dimension. We have imposed discrete timeslices, with one segment occupying each timeslice and coarticulation only operating over immediately adjacent timeslices. We have avoided the complexities of, for instance, smearing the elements over numerous, finer-grain timeslices (but see, for example, Gupta and Mozer 1993).

6.4. Statistical Modelling of the Corpus using a Connectionist Network

We have carried out extensive investigations of the statistical properties of both the phonetic and the feature-based versions of the corpus (see for instance Cairns *et al.* 1997). In the present chapter we present a complete picture of the acquisition of segmentation behaviour in these terms. We begin with the feature-based version of the corpus described above and a minimal processor that assumes only the nine different categories represented by the elements of Government Phonology. We progress from this simplest statistical model through processors that are increasingly complex, assuming more and more sophisticated levels of representation. At each stage we assess the implications for segmentation performance and we explore the relationship with the foregoing stage of the model in order to understand how the adult segmentation competence may emerge.

To begin with we ask what statistical regularities are apparent in a large corpus consisting of nine parallel continuous streams of binary elements, and what are the implications for segmentation. We have employed a connectionist network to compute these statistics. With this approach, there is the potential for the model to register regularities between elements within a single timeslice, and between continuous, and perhaps discontinuous, timeslices. Moreover, the model may choose to operate over its own idiosyncratic combination of lengths of dependency (pairs, triples of items, and so on), whereas the more traditional statistical approach which we also describe below concentrates exclusively on pairs or triples of consecutive items.

The network used in our research employed the Backpropagation-Through-Time (BPTT) algorithm (Rumelhart, Hinton, and Williams 1986) to learn the regularities of the corpus. Compared to the computationally more straightforward simple recurrent net, a network employing the BPTT algorithm allows the error signal to be backpropagated uncorrupted over more timeslices and was therefore judged to be preferable for the current task. The network received as input the binary vectors corresponding to the Government Phonology transcription, one per timeslice. At any one moment only the 'current' timeslice was visible to the model at its input. Sixty hidden units mediated between this input and the output units, at which the model was required to generate three vectors, corresponding respectively to the 'current', the 'previous' and the 'next' timeslices. The requirement to remember a past timeslice and predict the next one forced the network to learn the regularities contained in the corpus, as did the imposition of limited random noise (flipping binary features with a certain probability). The noise meant that the network could be less sure of the precise contents of any one timeslice and was thereby forced to rely more on the immediate context to adjudicate the identity of a segment compromised by noise. A cross entropy error measure was used (Hinton 1989), with training consisting of two passes over a section of the corpus 1 million segments long, and with the learning rate being reduced as training progressed.

The statistical approach to segmentation which we explore here means that we are principally concerned with the error at the *next* timeslice at the output. The constraints present within a word or syllable should increase predictability, but if the current timeslice represents the last segment of a word, then it should be relatively difficult to predict the contents of the next timeslice and the error should be high. Thus, in a continuous stream of error measurements for the prediction of the next timeslice at each point, the peaks in the stream should tend to correspond to word boundaries. There is no one measurement of the results of this approach; the outcome will depend on where the cut-off is placed. If the initial assumption of unpredictability at word boundaries has any value, then a very high cut-off point will be a conservative measure which settles for hitting fewer correct boundaries but is misled into fewer false-alarms (declaring a boundary incorrectly); conversely, a more liberal criterion will guarantee that more correct boundaries are hit, but at the expense of more false-alarms. The ROC graph in Figure 6.1 shows the network's performance, with a hit rate and

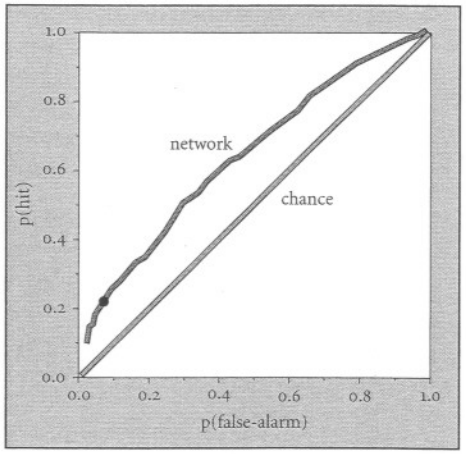


FIGURE 6.1. ROC graph showing the network's segmentation performance, adapted from Cairns *et al.* (1997)

false-alarm rate varying over different cut-off points. Greatest success is represented by a curve that departs most from the chance level represented by the diagonal. The strictness of the criterion used may vary depending on a range of other factors, which can only be assessed in the context of a fuller theory. For instance, over-segmenting (generating more false-alarms) may lead the infant to store too many (partial) 'words', whereas under-segmenting may result in too few (compound) 'words' being stored. These two outcomes will have different consequences. What we emphasize with the use of the ROC graph is the range of segmentation performance that a particular type of information can provide; precisely where on the graph represents the performance in any one case may be determined by other factors.

The assessment of a particular model's performance involves both statistical and psychological criteria. The dot on the curve in Figure 6.1 corresponds to the maximum value of the information theoretic measure *mutual information*, which we discuss elsewhere (Cairns *et al.* 1997):[2] at this point on the curve 21 per cent of the word boundaries are correctly identified, with a hits:false-alarm ratio of 1.5:1. However, it is misleading, and ultimately contrary to the approach we take here, to use the single statistic of number of boundaries detected to characterize any segmentation model that relies on a quantitative criterion, such as a cut-off point at a certain probability. As the ROC graph shows, hit rate can be improved at the expense of false-alarm rate. The optimal compromise between the two is a psychological question that involves model-specific issues of the cost of false-alarms relative to misses (failures to identify boundaries). For instance, there may be a different optimum point on the curve depending on the size and contents of the lexicon (if a lexicon exists at the stage considered), on the other segmentation algorithms available, or on the quality of the speech input available at the time. A high level of false-alarms may be relatively acceptable at a very early stage of development, when populating a lexicon with 'words' of a variable boundary-accuracy may be more important than trying to segment the speech stream veridically. Later on in development, false-alarms from this simplest sta-

[2] The mutual information of two sources, $M_{S,T}$ is defined as follows: $M_{S,T} = I_S + I_T - I_{S,T}$ where I_S and I_T are the total information of sources S with states s_i and T with states t_i respectively, and $I_{S,T}$ is the joint information between S and T. Thus:

$$I_S = -\sum_i p(s_i)\log(p(s_i))$$
$$I_T = -\sum_i p(t_i)\log(p(t_i))$$
$$I_{S,T} = -\sum_{ij} p(s_i, t_j)\log(p(s_i, t_j))$$

For the binary data with which we are concerned, each source has only two states (*boundary-present* and *boundary-absent*) yielding four possible combinations: *hit*, *false alarm*, *miss*, and *correct rejection*. The mutual information measure tests whether the general shape of the distributions of boundary points is the same for the segmentation algorithm and the veridically segmented corpus, and the extent to which the individual decisions match.

tistical strategy may be less important if they can be offset against different distributions of false-alarms from coexisting strategies. Finally, we may speculate that false-alarms may even have some *positive* features associated with them in that they may reveal to the lexicon the morphological structure of complex words.

The segmentation performance of this simplest model, whatever the cut-off point, is modest in many respects, although it was significantly better than a random segmentation algorithm that was designed to produce the same number of boundaries with 'word'-lengths of comparable distribution. Further, analysis of the network's false-alarms showed that they were respecting English syllable structure; in effect, the network was not distinguishing word boundaries from syllable boundaries, which is quite appropriate given the preponderance of monosyllables in conversational English.

In isolation, the model's performance represents only the first foothold on the segmentation problem. Note that a hit rate of 21 per cent does not represent 21 per cent of words isolated; a single word will only be isolated when two such boundaries border the same word. If the output of this segmentation process is stored, then the resulting 'lexicon' will mostly contain short strings of two, three, four, or more words, sometimes starting or finishing midway through a word, and occasionally representing a single isolated word. Whereas we may speculate that even these fragments might be consolidated into a lexicon of single words, given a sufficient number of overlapping fragments being laid down in the appropriate architecture, the performance of this minimal model is actually considerably better than it first appears, as it may be augmented by those boundaries defined by pauses and changes of speaker. (Pausing is the only other inarguably bottom-up evidence for a boundary, and has in fact been proposed as a means of getting the lexicon started (Suomi 1993).) This additional information does not add to the representational complexity of the model. When the 21 per cent of boundaries recognized in Figure 6.1 are combined with those revealed by pausing, the total is 32 per cent, the hits:false-alarms ratio being around 3:1. Almost one third of word boundaries may be identified by a processor which is sensitive to local, featural statistical regularities, including pausing; again, this figure is only illustrative—false-alarm considerations may increase or decrease it. This performance represents a definite first step—and a bottom up one at that—on the path to adult segmentation competence.

During the first six months of life the infant begins to perceive speech sounds categorically (see for example Kuhl *et al.* 1992), and this structuring of an internal phonological space continues past the first year (see Werker 1993 for a review of the evidence). If we increase the sophistication of our model by allowing it a phonemic level of representation, then we considerably improve its capacity to segment the speech stream. We constructed *n*-gram models of segmentation by counting the frequency of occurrence of all sequences of *n* contiguous items throughout the corpus and then hypothesizing word boundaries at points where the *n*-gram frequency is low. Figure 6.2 shows the ROC graph for three *n*-gram

measures of segmentation, using (non-connectionist) probability statistics. In this approach, it was hypothesized that the lower a bigram's or trigram's frequency of occurrence the more likely it is to straddle a word-boundary. As the different phonemic categories are only being established during the first half of the first year of infancy, all we assume is the simple ability to distinguish between these categories (perhaps even on the basis of diphone storage in the case of the bigrams). The relevant performance in the ROC graph in Figure 6.2 is the curve for the simple bigrams, which shows an improvement over the performance of the network model. The introduction of a phonetic level and explicit *n*-gram statistics have together provided the model with a better strategy. Note that the increasingly powerful and complex sources of information being incorporated into the model do not necessarily supplant the earlier sources of information; together they represent a more and more effective conspiracy of segmentation cues, allowing the developing processor to identify a wider range of word boundaries.

The bigrams strategy referred to in Figure 6.2 required only that the different segment identities be distinguished, irrespective of frequency; a particular error score was treated identically when predicting both the frequent segment /ə/ and the much less frequent segment /ʒ/. However, as the first year of life continues, more and more experience with these phonetic categories accrues; the brain gathers information about the frequency of the different segments and their distributions. This information may be used to construct a yet more sophisticated model, one in which the probability of occurrence of a particular bigram is *normalized* by taking into account the probability of occurrence of its constituent segments. It is precisely this information that produces the best perfor-

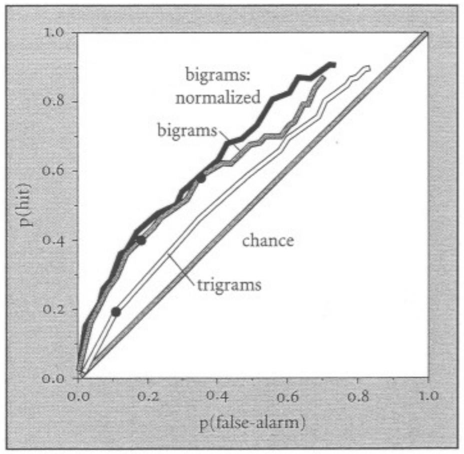


FIGURE 6.2. ROC graph showing segmentation performance using traditional *n*-grams

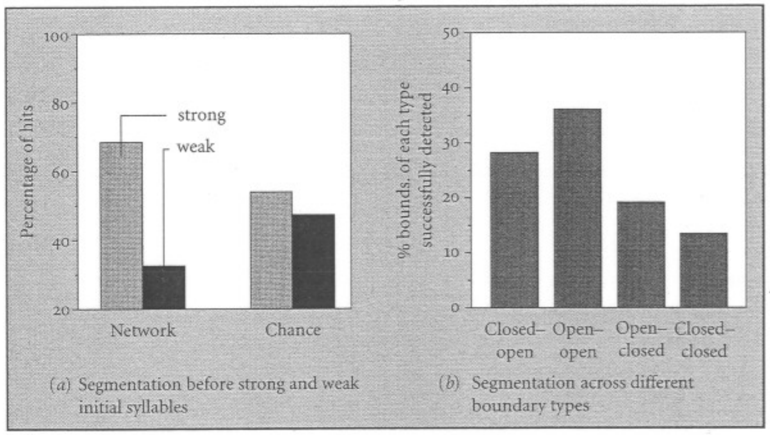
mance in Figure 6.2 when the bigram measure is normalized for phoneme frequency. This new model gives, for instance, a success rate of 38 per cent with a hits:false-alarm ratio 0.85:1.

The previous figure of 38 per cent is almost identical to the Harrington *et al.* trigram-model performance, referred to above; in fact, the latter was more reliable, having a hits:false-alarms ratio of around 8:1. It should be noted that the information assumed in the current bigram and trigram approach is radically less sophisticated than that assumed in the Harrington *et al.* study. The results in Figure 6.2 were obtained without reference to the actual identity of the word boundaries, and hence still represent a 'strongly bottom-up approach', whereas the approach used by Harrington *et al.* made use of a dictionary to obtain known word-boundary information and hence had recourse to lexical knowledge. We will not assume a lexical level of representation until later in the progression of models we are currently describing.

We move on now to the next most advanced model and to the segmentation algorithm that has attracted most of the recent attention in studies of adult segmentation performance in English: the Metrical Segmentation Strategy (MSS) (Cutler and Norris 1988). This strategy is seen by Cutler *et al.* as the instantiation, for spoken English, of a universal strategy to segment speech according to prosodic criteria. The strategy requires the processor to posit a word boundary before any strong syllable,[3] where a 'strong syllable' is defined as one bearing primary or secondary stress and containing a full vowel. Strong syllables contrast with weak syllables, which are unstressed and contain short, central vowels such as /ə/. Jusczyk, Cutler, and Redanz (1993) have shown that the MSS develops somewhere between the ages of six and nine months in infants in an English-speaking environment. It therefore develops in parallel with the refining of the categorical perception of phonemes. The potential contribution of the MSS to lexical segmentation has been placed, at its most optimistic, at approximately 90 per cent of open-class words, in that this proportion of open-class words begins with a strong syllable (Cutler and Carter 1987). Note that more than half of the words in the LLC are closed-class words, two-thirds of which are probably realized as weak syllables. Our own calculations made on the basis of the distribution of word types in the LLC and on careful listening to a different, very much smaller, taped corpus of conversational English, show that strong syllables, on which the MSS depends, potentially identify some 50 per cent of all word boundaries, with an almost negligible false-alarm rate (that is, non-word-initial strong syllables).[4] The MSS can clearly contribute very substantially to segmentation behaviour. However, it must be bought at the cost of developing a categorical distinction between strong and weak syllables, meaning that our model becomes even more sophisticated as it now possesses a prosodic level of representation in addition to its existing phonetic competence.

In Figure 6.3(*a*) we see that the emergence of the MSS is prefigured in the segmentation performance of the network, our simplest model which operates on information taken to be available from earliest infancy. The network prefers to segment the speech stream before strong syllables. This does not constitute a reduction of the MSS to the level of feature-based processing. Rather, it allows us to conceive of how the MSS might emerge from the general statistical extraction of regularities from the speech stream, rather than, for instance, by some 'parameter-setting' on a hand-wired, limited choice between metrical, syllabic, and moraic strategies. It is part of the brain's characteristic activity that it is able to create discrete processing modules geared to specific input regularities, which may then interact very flexibly with the rest of processing. However, the input regularities must first become apparent to the processor to prompt the development of categorical distinctions, in this case the distinction between strong and weak syllables. Note finally that when the network's calculations were normalized for segment frequency (by dividing each prediction error score by the frequency of the segment being predicted) the network no longer showed a preference for segmenting before strong syllables and before open-class words, and showed instead a preference for detecting initial closed-class boundaries.

Each increasingly sophisticated level of competence proves to be more powerful in its segmentation ability than the previous competence. The MSS is the most powerful strategy encountered so far in this account of the development



(*a*) Segmentation before strong and weak initial syllables

(*b*) Segmentation across different boundary types

FIGURE 6.3. Network performance in relation to the MSS and to syntactic variables

[3] In later work by Cutler *et al.*, the necessity of inserting a word boundary is changed to a quantitative contribution of metrical prosody to the competition between simultaneously activated lexical hypotheses.

[4] We gloss over the fact that the MSS necessitates a syllabification strategy that may be non-trivial for complex clusters of consonants.

of segmentation ability. The 50 per cent of boundaries which we calculate the MSS potentially identifies are overwhelmingly the onsets of open-class words. Figure 6.3(*b*) shows that the network model identifies a proportion of the onsets of closed-class words. Combining the success of the MSS with the performance of simpler strategies, the segmentation problem seems to be much less daunting. The conspiracy of cues at this stage will correctly identify the majority of word boundaries. The resulting lexicon will increasingly be supplied with single words and those badly segmented 'words' it contains from the earlier stages will find no confirmation. A lexical competence emerges.

With the appearance of reliable lexical categories, the stage is set for the most powerful segmentation information that we have seen in this developmental progression. Now the processor is able to compile reliable information concerning the probabilities that individual bigrams and trigrams straddle word boundaries. Note that the processor will have been able to start doing this as soon as phonetic identities were established, although its 'words' will have been far less reliably defined. We see a spiralling reliability in defining segments and words which gives rise to ever-increasing segmentation performance. Once categorical phonotactic information is assumed, then it is possible to investigate the utility of specific types of information at that level by supplying the model with the relevant knowledge, such as, for instance, all the legal initial and final consonant clusters, or syllable-internal constraints reflecting the sonority hierarchy; see Cartwright and Brent (1994); Brent and Cartwright (1996).

Figure 6.4 shows the powerful segmentation capacity of *n*-gram models with reliable lexical level representations. These curves were obtained by calculating the relative probabilities that particular *n*-grams did or did not straddle a word boundary in the LLC. A cut-off point is then established, so that a boundary is hypothesized if its probability is sufficiently small given a particular *n*-gram. The most conservative cut-off point for bigrams gives a segmentation performance that converges on that obtained by Harrington *et al.* (who employed an all-or-none approach as opposed to a probabilistic one); cut-off points that allow for a measure of false-alarms give extremely good segmentation performance. For instance, the mutual information maximum for the bigrams in this model gives a 75 per cent detection rate with a hits:false-alarms ratio of 4.7:1. Performance is even more powerful when the statistics are calculated over trigrams: for instance, 93 per cent of boundaries detected at the mutual information maximum, with a hits:false-alarm ratio of 9:1. It might be argued that this powerful performance is premised on a correct close phonetic transcription, information that is not always available in the perception of conversational speech. Accordingly, we retranscribed the corpus into six broad phonetic classes, as in the approach taken by Zue and his colleagues (for example, Huttenlocher and Zue 1983): stops, nasals, weak fricatives, strong fricatives, liquids and glides, vowels. Even without finegrain phonological information the segmentation performance using a bigram measure is 74 per cent detection, with a 1.5:1 hits:false-alarms ratio at the mutual information maximum, a performance which compares favourably with those of the models encountered so far. In summary, then, the initial development of a lexical competence dramatically increases segmentation performance in that it allows the calculation of veridical bigram and trigram probabilities across word boundaries.

At this point in our progression through models of increasing complexity, we have shown that sufficient structure may be discovered in the speech stream, with minimal initial assumptions, to make substantial inroads into the segmentation problem. Not only do the models at this latest level of competence perform better than any individual previous ones, but, overall, a *conspiracy* of the strategies we have considered now seems to be enough to overwhelm the problem. Nevertheless, we will continue with our consideration of models of increasing sophistication. Before leaving those models with a lexical-level competence, it should be remarked that a novel element appears at this stage. For the first time it is possible for the processor to *predict* when a word will end. The lexicon contains stored phonological representations of words and if the beginning of one of these matches the current speech stream, then it is possible to predict the boundary at which this word ends and the next starts. Until this stage in the progression it has not been possible to make useful boundary predictions other than those concerning the next timeslice. The utility of such predictions is affected by several factors, however. First, Luce (1986) has calculated that a high percentage of short words are not phonologically unique by their offsets, meaning that it will often not be feasible to predict a word ending reliably as the

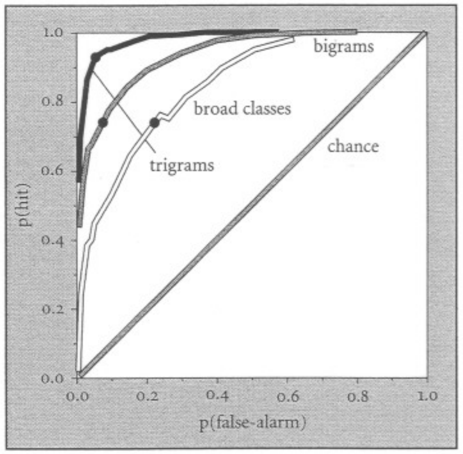


FIGURE 6.4. ROC graph for *n*-gram performance when veridical knowledge of *n*-gram probabilities across word boundaries is known

current cohort of lexical candidates will contain more than one word. This observation is particularly relevant given that conversational English is dominated by short words; the mean length of the words in the LLC is 3.7 segments. Second, Luce's dictionary-based study does not take into account the unpredictability of closed-class suffixes: knowing that *park* is a word will not predict the end of *parks* or *parking*. However, when knowledge of the phonological form of words is augmented by their frequencies, predictability is probably increased considerably, given that lexical cohorts typically consist of only one or two high-frequency members and a larger number of low-frequency members (Bard and Shillcock 1993).

The next level of competence to be achieved concerns syntactic processing. Shillcock and Bard (1993) have demonstrated that there is genuine 'top-down' interaction between syntactic processing and the lexical-level representations of closed-class words. We may speculate that this behaviour associated with closed-class words also holds for bound morphemes, the suffixes which are unpredictable on the basis of lexical information alone. The development of syntactic competence will allow predictions about the identity of the large proportion of closed-class words in conversational speech, and the location of their boundaries. Figure 6.3(*b*) shows that the segmentation behaviour of the simplest network model that we consider is not irrelevant to the syntactic dimension. There is a clear gradation of segmentation performance, with boundaries associated with open-class words being more likely to be identified: performance is best with open–open boundaries and worst with closed–closed boundaries. From the earliest segmentation behaviour, then, there is a preference for isolating open-class words in English. When we assess the individual words completely isolated by the network (that is, both boundaries identified) there are significantly more open-class words than chance: 59 per cent of those extracted, compared with 40 per cent in the corpus as a whole. This concentration on open-class words is accentuated by the development of the MSS (which is disproportionately more successful with open-class word beginnings), but is counteracted by the development of accurate segment frequency information, which allows normalized probabilities to be calculated, thus offsetting the tendency for those (frequent) segments contained in the very high frequency closed-class words to be very easy to predict.

The final competences associated with semantic and pragmatic processing simply add to the effectiveness of the predictions of the ends of words, and we will not discuss these competences further.

We have seen that a statistical analysis of the 'phonological space' being developed by the infant gives an insightful account of the development of segmentation ability. In particular, connectionist modelling involving a Government Phonology-based representation of phonological space shows how a processor with minimal representational assumptions can gain an initial foothold on the segmentation problem and then step up into ever more complex processing by making more and more representational innovations.

Our analysis of the development of segmentation behaviour, particularly its early stages, would be stronger if we could demonstrate that the connectionist modelling we have described can provide an account of other language processing phenomena, unrelated to segmentation. Below, we briefly refer to two other phenomena.

First, our model gives a parsimonious account of a widely discussed demonstration of phoneme restoration in a 'compensation for coarticulation effect'. Elman and McClelland (1988) present data from an elegant experiment which they claim is evidence for 'the cognitive penetration of the mechanisms of perception', or genuine top-down processing. Using the connectionist model, the spoken language corpus and the phonological representations we have described above, we have demonstrated that the results reported by Elman and McClelland are also explained as purely bottom-up effects based on the phonological statistics of spoken English (Shillcock *et al.* 1993).

Second, initial explorations suggest our approach to modelling phonological space can capture some of the data concerning the acquisition profile of individual phonological segments (Shillcock *et al.* 1993) as reported by Sander (1972) and Prather, Hedrick, and Kern (1975). By interrupting the training of the network and testing it at regular intervals to see on which segments a critical performance has been achieved, we can obtain a profile to compare with the human developmental data. Our initial observations show that the sequence of segments obtained from the model is marginally significantly correlated ($p < .1$, using Kendall's *tau*) with the human data. In addition, the early establishment of vowel representations mirrors human performance (Kuhl *et al.* 1992). The order of acquisition by the model does not simply reflect the overall frequency of the different segments in the training corpus.

6.5. Discussion and Conclusions

Connectionist statistical modelling has allowed us to assess the possibility of acquiring effective segmentation behaviour with minimal assumptions as to the initial state of the processor. All that is assumed is the brain's sensitivity to the small-scale statistical structure of the stimulus environment. Perhaps we might view the problem not from the point of view of where the speech stream might best be segmented (this assumes that active segmentation should occur) but where a representation of the speech stream will naturally 'snap' if the only rule is 'try and store the entire speech stream (bounded, perhaps, by pauses)'. This strategy is clearly impossible—the attempted representations will disintegrate into smaller ones—but what we have shown is that, if these disintegrations occur at points of low predictability, then there is a clear tendency for them to occur at word boundaries. We have seen that the network's predictions, together with pausing, can account for 32 per cent of word boundaries (with around a 3:1

current cohort of lexical candidates will contain more than one word. This observation is particularly relevant given that conversational English is dominated by short words; the mean length of the words in the LLC is 3.7 segments. Second, Luce's dictionary-based study does not take into account the unpredictability of closed-class suffixes: knowing that *park* is a word will not predict the end of *parks* or *parking*. However, when knowledge of the phonological form of words is augmented by their frequencies, predictability is probably increased considerably, given that lexical cohorts typically consist of only one or two high-frequency members and a larger number of low-frequency members (Bard and Shillcock 1993).

The next level of competence to be achieved concerns syntactic processing. Shillcock and Bard (1993) have demonstrated that there is genuine 'top-down' interaction between syntactic processing and the lexical-level representations of closed-class words. We may speculate that this behaviour associated with closed-class words also holds for bound morphemes, the suffixes which are unpredictable on the basis of lexical information alone. The development of syntactic competence will allow predictions about the identity of the large proportion of closed-class words in conversational speech, and the location of their boundaries. Figure 6.3(*b*) shows that the segmentation behaviour of the simplest network model that we consider is not irrelevant to the syntactic dimension. There is a clear gradation of segmentation performance, with boundaries associated with open-class words being more likely to be identified: performance is best with open–open boundaries and worst with closed–closed boundaries. From the earliest segmentation behaviour, then, there is a preference for isolating open-class words in English. When we assess the individual words completely isolated by the network (that is, both boundaries identified) there are significantly more open-class words than chance: 59 per cent of those extracted, compared with 40 per cent in the corpus as a whole. This concentration on open-class words is accentuated by the development of the MSS (which is disproportionately more successful with open-class word beginnings), but is counteracted by the development of accurate segment frequency information, which allows normalized probabilities to be calculated, thus offsetting the tendency for those (frequent) segments contained in the very high frequency closed-class words to be very easy to predict.

The final competences associated with semantic and pragmatic processing simply add to the effectiveness of the predictions of the ends of words, and we will not discuss these competences further.

We have seen that a statistical analysis of the 'phonological space' being developed by the infant gives an insightful account of the development of segmentation ability. In particular, connectionist modelling involving a Government Phonology-based representation of phonological space shows how a processor with minimal representational assumptions can gain an initial foothold on the segmentation problem and then step up into ever more complex processing by making more and more representational innovations.

Our analysis of the development of segmentation behaviour, particularly its early stages, would be stronger if we could demonstrate that the connectionist modelling we have described can provide an account of other language processing phenomena, unrelated to segmentation. Below, we briefly refer to two other phenomena.

First, our model gives a parsimonious account of a widely discussed demonstration of phoneme restoration in a 'compensation for coarticulation effect'. Elman and McClelland (1988) present data from an elegant experiment which they claim is evidence for 'the cognitive penetration of the mechanisms of perception', or genuine top-down processing. Using the connectionist model, the spoken language corpus and the phonological representations we have described above, we have demonstrated that the results reported by Elman and McClelland are also explained as purely bottom-up effects based on the phonological statistics of spoken English (Shillcock *et al.* 1993).

Second, initial explorations suggest our approach to modelling phonological space can capture some of the data concerning the acquisition profile of individual phonological segments (Shillcock *et al.* 1993) as reported by Sander (1972) and Prather, Hedrick, and Kern (1975). By interrupting the training of the network and testing it at regular intervals to see on which segments a critical performance has been achieved, we can obtain a profile to compare with the human developmental data. Our initial observations show that the sequence of segments obtained from the model is marginally significantly correlated ($p < .1$, using Kendall's *tau*) with the human data. In addition, the early establishment of vowel representations mirrors human performance (Kuhl *et al.* 1992). The order of acquisition by the model does not simply reflect the overall frequency of the different segments in the training corpus.

6.5. Discussion and Conclusions

Connectionist statistical modelling has allowed us to assess the possibility of acquiring effective segmentation behaviour with minimal assumptions as to the initial state of the processor. All that is assumed is the brain's sensitivity to the small-scale statistical structure of the stimulus environment. Perhaps we might view the problem not from the point of view of where the speech stream might best be segmented (this assumes that active segmentation should occur) but where a representation of the speech stream will naturally 'snap' if the only rule is 'try and store the entire speech stream (bounded, perhaps, by pauses)'. This strategy is clearly impossible—the attempted representations will disintegrate into smaller ones—but what we have shown is that, if these disintegrations occur at points of low predictability, then there is a clear tendency for them to occur at word boundaries. We have seen that the network's predictions, together with pausing, can account for 32 per cent of word boundaries (with around a 3:1

hits:false-alarm rate). As a result of the low-level statistics of the spoken English speech stream, the infant is set on a course of acquiring a word-based lexicon. The network's performance was modest and a brief comparison with the *n*-gram results suggests that its performance was overwhelmingly determined by relationships spanning only two or three timeslices. Nevertheless, its strengths lie in the fact that it assumes very little (not even an active predisposition to segment) and that it opens the way for the discovery of potentially more effective segmentation cues such as the Metrical Segmentation Strategy.

Throughout, we have simply assumed that there are real-world interactions occurring, between the infant and others, to consolidate the lexical entries that the different statistical strategies suggest, providing referential content and potentially allowing erroneously stored multiword strings to be decomposed.

The eventual success of the adult listener in segmenting speech may be due to the ability to recruit results flexibly from all of the levels of segmentation competence we have described, basing segmentation judgments on the sensitivity of the different competences to different aspects of the input. Thus, for instance, at the very beginning of an utterance higher-level competences may be relatively less useful, as little lexical, syntactic, or semantic context has been established. At this point, rhythmic or low-level cues may be particularly important. The ability of the adult processor to respond flexibly is shown by the fact that even the 'best' French–English bilinguals possess only one prosodic segmentation strategy (metrical or syllabic) (Cutler *et al.* 1992), indicating perhaps that the flawed results of an inappropriate prosodic segmentation strategy might be offset by greater reliance on the rest of the range of cues. Similarly, when the speech quality is poor, segmentation strategies based on full knowledge of phonological features are untenable, and processing will rely more heavily on the metrical strategy and on the level of performance possible with only 'broad class' transcriptions. At those points in the speech stream when segmentation is heavily determined by higher-level contextual factors, then the low-level statistical processing we have described will be relatively superfluous.

The simplest overall model of the processor would perhaps ascribe each of the levels of processing we have considered to a separate module, allowing the outputs of the individual modules to contribute to an overall boundary likelihood at any one point in time. The value of flexibility in the importance attached to the different sources of information is shown by the fact that each source possessed its own strengths and weaknesses. Thus, for instance, when the network results and the bigram results were normalized for segment frequency, the preponderance in the identification of strong-syllable boundaries and open-class words ceased, and, in the closed–open case, a reversal occurred in which more closed-class word-boundaries were identified. In contrast, the MSS has an inherent bias towards detecting open-class word-beginnings.

In conclusion, our connectionist and non-connectionist statistical approach provides a psychologically plausible account of the different types of information

that the developing processor can draw upon to solve the problem of speech segmentation and perform flexibly in the adult state. The segmentation results obtained demonstrate that the relevant information is there in the speech stream, awaiting discovery by the developing infant.

References

Altmann, G. T. M. and Shillcock, R. C. (eds.) (1993). *Cognitive Models of Speech Processing. The Second Sperlonga Meeting.* Hillsdale, NJ: Erlbaum.

Bard, E. G. and Shillcock, R. C. (1993). 'Competitor Effects During Lexical Access: Chasing Zipf's Tail'. In Altmann and Shillcock (eds.), 235–75.

Brent, M. R. (1993). 'Minimal Generative Explanations: A Middle Ground Between Neurons and Triggers'. *Proceedings of the 15th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum, 30–6.

—— and Cartwright, T. A. (1996). 'Distributional Regularity and Phonotactic Constraints Are Useful for Segmentation'. *Cognition*, 61, 93–125.

Briscoe, E. J. (1989). 'Lexical Access in Connected Speech Recognition'. *Proceedings of the 27th Congress, Association for Computational Linguistics,* Vancouver, 84–90.

Cairns, P., Shillcock, R. C., Chater, N., and Levy, J. (1997). 'Bootstrapping Word Boundaries: A Bottom-up Corpus-based Approach to Speech Segmentation'. *Cognitive Psychology*, 33, 111–53.

Cartwright, T. A. and Brent, M. R. (1994). 'Segmenting Speech Without a Lexicon: Evidence for a Bootstrapping Model of Lexical Acquisition'. *Proceedings of the 16th Annual Conference of the Cognitive Sience Society.* Hillsdale, NJ: Erlbaum, 148–52.

Cole, R. A. and Jakimik, J. (1980). 'A Model of Speech Perception'. In R. A. Cole (ed.), *Perception and Production of Fluent Speech.* Hillsdale, NJ: Erlbaum.

Cutler, A. and Butterfield, S. (1992). 'Rhythmic Cues to Speech Segmentation: Evidence from Juncture Misperception'. *Journal of Memory and Language*, 31, 218–36.

—— and Carter, D. M. (1987). 'The Predominance of Strong Initial Syllables in the English Vocabulary'. *Computer Speech and Language*, 2, 133–42.

—— and Norris, D. (1988). 'The Role of Strong Syllables in Segmentation for Lexical Access'. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–21.

—— Mehler, J., Norris, D., and Segui, J. (1992). 'The Monolingual Nature of Speech Segmentation by Bilinguals'. *Cognitive Psychology*, 24, 381–410.

Elman, J. L. and McClelland, J. L. (1988). 'Cognitive Penetration of the Mechanisms of Perception: Compensation for Coarticulation of Lexically Restored Phonemes'. *Journal of Memory and Language*, 27, 143–65.

Gupta, P. and Mozer, M. C. (1993). 'The Nature and Development of Phonological Representations: Network Explorations'. *Proceedings of the 15th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum, 516–21.

Harrington, J., Watson, G., and Cooper, M. (1988). 'Word Boundary Identification from Phoneme Sequence Constraints in Automatic Continuous Speech Recognition'. *Twelfth International Conference on Computational Linguistics, Coling '88*, Budapest, August 1988.

HARRINGTON, J., WATSON, G., and COOPER, M. (1989). 'Word Boundary Detection in Broad Class and Phoneme Strings'. *Computer Speech and Language*, **3**, 367–82.

HARRIS, J. and LINDSEY, G. (1993). 'The Elements of Phonological Representation'. In J. Durand and F. Katamba (eds.), *New Frontiers in Phonology*. Harlow, UK: Longman.

HINTON, G. E. (1989). 'Connectionist Learning Procedures'. *Artificial Intelligence*, **40**, 185–234.

HUTTENLOCHER, D. P. and ZUE, V. W. (1983). 'Phonotactic and Lexical Constraints in Speech Recognition'. *Proceedings of the 3rd National Conference on Artificial Intelligence*. AAAI Press, California, 172–6.

JUSCZYK, P. W., CUTLER, A., and REDANZ, N. (1993). 'Infants' Sensitivity to Predominant Word Stress in English'. *Child Development*, **64**, 675–87.

KAYE, J. D., LOWENSTAMM, J., and VERGNAUD, J.-R. (1985). 'The Internal Structure of Phonological Elements: A Theory of Charm and Government'. *Phonology Yearbook* **2**, 305–28.

KUHL, P., WILLIAMS, K. A., LACERDA, F., STEVENS, K. N., and LINDBLOM, B. (1992). 'Linguistic Experience Alters Phonetic Perception in Infants by Six Months of Age'. *Science*, **255**, 606–8.

LUCE, P. (1986). 'The Computational Analysis of Uniqueness Points in Auditory Word Recognition'. *Perception and Psychophysics*, **39**, 155–8.

PRATHER, E. M., HEDRICK, D. L., and KERN, C. A. (1975). 'Articulation Development in Children Aged Two to Four Years'. *Journal of Speech and Hearing Disorders*, **20**, 179–91.

REDLICH, A. N. (1993). 'Redundancy Reduction as a Strategy for Unsupervised Learning'. *Neural Computation*, **5**, 289–304.

RUMELHART, D. E., HINTON, G. E., and WILLIAMS, R. J. (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*. Cambridge, Mass: MIT Press, 318–62.

SANDER, K. (1972). 'When are Speech Sounds Learned?' *Journal of Speech and Hearing Disorders*, **37**, 55–63.

SHILLCOCK, R. C. and BARD, E. G. (1993). 'Modularity and the Processing of Closed Class Words'. In Altmann and Shillcock (eds.), 163–85.

—— LEVY, J. P., LINDSEY, G., CAIRNS, P., and CHATER, N. (1993). 'Connectionist Modelling of Phonological Space'. In T. M. Ellison and J. M. Scobbie (eds.), *Computational Phonology. Edinburgh Working Papers in Cognitive Science*, **8**, 179–95.

—— LINDSEY, G., LEVY, J., and CHATER, N. (1992). 'A Phonologically Motivated Input Representation for the Modelling of Auditory Word Perception in Continuous Speech'. *Proceedings of the 14th Annual Cognitive Science Society Conference*, Bloomington, 408–13.

SUOMI, K. (1993). 'An Outline of a Developmental Model of Adult Phonological Organization and Behaviour'. *Journal of Phonetics*, **21**, 29–60.

SVARTVIK, J. and QUIRK, R. (1980). *A Corpus of English Conversation*. Lund: Gleerup.

WERKER, J. (1993). 'Developmental Changes in Cross-language Speech Perception: Implications For Cognitive Models of Speech Processing'. In Altmann and Shillcock (eds.), 57–78.