The Production and Perception of Word Boundaries

Anne Cutler

Applied Psychology Unit
Medical Research Council
Cambridge, UK

Introduction: The Word Boundary Problem

The recognition of continuous speech presents listeners (human or machine) with a problem which does not arise in the recognition of isolated words, and does not confront the reader in most orthographies. The act of recognition is the identification of an input as something we already know; what we already know is not the whole of an input utterance, because human memory is not infinite, and it would be impossible to store in our memories every complete utterance we might ever hear. Therefore the entries in our mental lexicon must be *discrete*, and recognition will involve finding these discrete lexical units as sound patterns in the speech signal input, and matching them to lexical entries in order to determine their meaning. The problem for listeners arises in the fact that speech is continuous: lexical unit boundaries are not reliably marked. Finding the boundaries between such units — *segmenting* the speech signal — is therefore a non-trivial task for all listeners.

In our laboratory we have studied the word boundary problem from several different perspectives; this report presents only a brief summary of each line of research.

Strategies for Prelexical Segmentation

Listeners respond to the challenge of the word boundary problem by developing segmentation heuristics based on their knowledge of linguistic

regularities. Evidence from our experiments supports a strategy for English which we call "metrical segmentation". The strategy exploits the characteristic rhythm of English speech, i.e., the opposition of strong versus weak syllables. Strong syllables are those which contain full vowels; weak syllables are those which contain central, or "reduced" vowels. Consider the four words *generous, generic, generate* and *generation;* their strong-weak patterns are SWW, WSW, SWS and SWSW respectively. It can be seen that stressed syllables are necessarily strong, and weak syllables are necessarily unstressed; but levels of stress are not relevant to this binary distinction. Thus it is irrelevant that primary stress falls on the first syllable in *generate* and on the third in *generation;* the first and third syllable are strong in both cases.

The metrical segmentation strategy on which English-speaking listeners rely is: segment speech at the onset of all strong syllables. In other words, listeners treat any strong syllable as if it were highly likely to be word-initial.

The evidence for this postulated strategy comes both from laboratory studies and from naturalistic observation. Firstly, Cutler and Norris (1988) asked listeners to perform a task called "word spotting", which consisted of deciding whether or not a nonsense bisyllable began with a real word. They found that a word like *mint* is harder to detect in *mintayf* (two strong syllables) than in *mintef* (a strong and a weak syllable). They explained this result by suggesting that the second strong syllable in *mintayf* triggers segmentation, so that detection of the embedded word requires assembly of speech material across a point at which speech has been segmented.

Cutler and Butterfield (1991a) studied the pattern of errors which listeners make when they misperceive word boundaries in continuous speech. In spontaneous slips of the ear, they found, listeners significantly more often mistakenly insert word boundaries before strong syllables than before weak syllables; mistaken deletions of word boundaries, on the other hand, occur significantly more often before weak syllables than before strong. For example, the misperception of *shell officially* as *Sheila Fishley* involves deletion of the boundary before the second syllable of the utterance, and insertion of a boundary before the third syllable instead.

In an experiment in which they presented listeners with very faint speech, Cutler and Butterfield elicited just the same pattern of mistakenly inserting word boundaries before strong syllables, but mistakenly deleting word boundaries before weak syllables (e.g., *conduct ascents uphill* was reported as *the doctor sends her bill).*

Thus the strategy of segmenting English at strong syllable onsets is well entrenched in listener performance. Cutler and Carter (1987) pointed out that the strategy would in fact be a highly efficient way of segmenting English. The majority of lexical words (nouns, verbs and adjectives) in the English vocabulary do indeed begin with strong syllables; moreover, the

average frequency of occurrence of lexical words beginning with weak syllables is quite low, which further increases the likely proportion of strong initial syllables in typical speech contexts. Cutler and Carter analysed a 190,000 word corpus of spontaneous British English speech; they found that less than 10% of lexical words in this corpus began with weak syllables. Thus even the simplest form of metrical segmentation would correctly locate over 90% of lexical word onsets; this success rate is presumably high enough to make this heuristic very useful in continuous speech recognition.

Production of Word-Boundary Cues

Our second approach to the word boundary problem has been to study the production of word boundary cues. As noted above, such cues do not reliably occur in normal speech. But sometimes speakers become aware that listeners are having difficulty — whether because of background noise, imperfect linguistic ability, or some other reason. Under such conditions speakers usually try to speak particularly clearly. Given how important a task lexical segmentation is for the listener, it would seem that explicit word boundary cues would be a really useful aid. Therefore Cutler and Butterfield (1990a, 1991b; Butterfield and Cutler, 1990) elicited deliberately clear speech, in an attempt to study word boundary cues under conditions where they are most likely to occur. In the light of our previous work showing that listeners tend to treat strong syllables as word-initial but weak syllables as not word-initial, we were naturally interested in whether speakers would distinguish between types of word boundary, and mark some boundaries more than others.

In our experiments on clear speech, therefore, we constructed sentences containing particular word boundaries before strong *vs* weak syllables. We then had speakers produce these sentences, in the belief that a listener in the next room was trying to hear the sentences through a distorting filter. From each subject we first recorded baseline productions; we then gave the subject feedback suggesting each utterance had been misperceived in such a way that the crucial word boundary was omitted. Two further productions (in which the subject tried to speak more clearly) were then recorded, and we analysed the speech around the word boundary in the deliberately clear productions in comparison to the baseline.

The following is an example of the kind of sentences used in our experiments, and the "listener's" feedback responses (designed to be acceptable sentences, rhythmically and phonetically similar to what the subject said, but without the crucial word boundary):

Subject says: Take it in / turns to eat breakfast

Feedback 1: Baker interns all the terrorists
Feedback 2: Take it internally at breakfast

Subject says: He called in / to view it himself

Feedback 1: The cold interviewer was selfish
Feedback 2: He crawled into view by himself

Our analysis of the deliberately clear speech showed that speakers did indeed produce word boundary cues which were not present in the baseline. The principal way in which they signalled presence of a word boundary was by manipulating the durational pattern of the utterance — by pausing at word boundaries, and by lengthening pre-boundary syllables.

Interestingly, the boundary signals which speakers provided were significantly stronger at boundaries which preceded *weak* syllables (e.g., *in to)* than at boundaries preceding strong syllables (e.g., *in turns).* This suggests that speakers are allowing for the listener strategy of assuming that word boundaries are most likely to occur before strong syllables, and paying particular attention to marking those boundaries which this usual strategy would *not* detect. In this sense the way that speakers choose to make speech deliberately clear would appear to be very well adapted to listeners' needs.

Resolution of Boundary Ambiguities

In subsequent work we addressed the question of whether speakers' manipulations of their utterances in deliberately clear speech are actually of use to listeners. A necessary prerequisite is obviously that the capacity to exploit durational variation as a word boundary signal be part of the language user's competence. That a listener can readily exploit explicit pausing as a boundary marker presumably requires no special investigation; we therefore confined our attention to pre-boundary syllable lengthening. Previous work had not provided a clear answer to the question of whether such lengthening exists outside deliberately clear speech, and if so, whether listeners exploit it to perform lexical segmentation. Lehiste (1972) found no difference in the first syllables of bisyllabic strings such as *speeder* and *speed kills*, and concluded that "temporal readjustment processes tend to ignore ... word boundaries" (p. 2023). Likewise, Umeda (1975) found no significant difference in vowel durations for the same vowels occurring in monosyllables versus the stressed syllable of polysyllables. Beckman and Edwards (1990), however, reported word-final lengthening in juncturally ambiguous

strings such as "Pop opposed" versus "Poppa posed".

Such ambiguous sequences clearly offer a test case, and they were also used in a perceptual study by Taft (1984). She recorded bisyllables such as *lettuce* (which could also be *let us)*, or *invests* (which could also be *in vests)*, and had subjects judge whether they were one word or two. We adapted Taft's methodology to examine whether word-final lengthening occurs, and if it does occur, whether listeners exploit it in making segmentation decisions. In our study, a trained speaker produced 24 such ambiguous strings (12 strong-weak, 12 weak-strong) in final position in sentences (e.g., "I bought some apples and a beautiful lettuce"; "we'll go home early if our duties will let us"; "Every student in the form invests"; "All the children will be warm in vests"). We measured the durations of each syllable in the speaker's productions. We found that in weak-strong items both syllables were lengthened in the two-word version compared to the one-word, but in strong-weak items boundary-conditioned lengthening was confined to the first syllable. Even where there was lengthening, however, it was slight: the greatest amount of lengthening was 6.8%, in initial syllables of weak-strong items.

The ambiguous bisyllables were then excised from the sentence context and presented in isolation to listeners, who were asked to judge which context they had come from. We analysed both the correctness of their judgments, and the relationship between their judgments and the syllable durations.

The listeners showed a general preference for making one-word choices, presumably because items presented in isolation are more often expected to be one word than two. They also made significantly more two-word choices to two-word productions than to one-word productions. For weak-strong items (e.g., *inquires/in choirs)*, the proportion of two-word choices correlated positively with measured duration of both syllables. (Since first and second syllable durations were highly positively correlated, it is to be expected that each would show the same relationship to subject choice patterns.) For strong-weak items (e.g., *lettuce/let us)*, however, no correlations with any durational measure were statistically significant.

Nooteboom and Doodeman (1980) found that the just noticeable difference for vowel duration discriminating the Dutch words *tak* vs *taak* was about 5.5%. Except in the initial syllables of weak-strong items, the boundary-conditioned lengthening produced by our speaker was less than this. Thus it would appear that only in that one case did the amount of lengthening reach listeners' thresholds for durational discrimination. Our finding does suggest, however, that when durational cues to word boundaries are available, and are sufficient to exceed durational discrimination thresholds, listeners are able to exploit them.

Conclusions

Our studies of deliberately clear speech showed that speakers do try to give explicit cues to the presence of a word boundary. The cues which they use are durational: pausing at boundaries and lengthening of pre-boundary syllables. Our study of ambiguous bisyllables showed that listeners can make use of such temporal differences in deciding how to interpret the ambiguous sequence. This is evidence that speakers' clear speech strategies are indeed based on capacities commanded by listeners.

The differences, however, are small, and the effects they have on listener responses are also small. Our statistical studies of spontaneous speech showed that spoken English typically contains more monosyllables than polysyllables; thus it does not offer much scope for contrasting word-final with non-word-final syllables. Furthermore, the usefulness of word-final lengthening, as of any temporal cue, is limited by the fact that it can only be interpreted relative to the temporal pattern of the utterance in which it occurs. (In our ambiguous bisyllable study, the subjects' accuracy of identification — for weak-strong items — improved from the first to the second half of the perception experiment, presumably as they adjusted to the rate of speech.)

Our studies lead us to believe, as the previous literature had suggested, that English speech usually contains few temporal cues to word boundary location. To compensate for this, however, listeners have developed very efficient strategies for hypothesising where word boundaries are most likely to occur. Based as they are on the structure of the language itself, they work very well indeed. And perhaps the most significant of our findings is that speakers who are trying to speak deliberately clearly pay particular attention to marking boundaries before weak syllables. Our earlier work suggested that listeners use a "metrical segmentation strategy", which hypothesises that boundaries are most likely to occur before strong syllables. Thus speakers are taking care to mark just those boundaries which would not be detected by application of the usual listener strategies.

Acknowledgment

This paper presents a summary overview of the presentation given at the ATR Workshop on Speech Perception, Production and Linguistic Structure, November 1990. The presentation described work all of which is, or shortly will be, published elsewhere, and further experimental details may be obtained from the papers referred to in the text. The research described was carried out in collaboration with Sally Butterfield, David Carter and Dennis

Norris, and was financially supported by the Alvey Directorate, UK, IBM UK Scientific Centre, British Telecom and ESPRIT Basic Research Actions; to all of these, many thanks.

References

Beckman, M. & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. Beckman (Eds.) *Papers in laboratory phonology I: Between the grammar and the physics of speech.* Cambridge: Cambridge University Press; pp. 152-178.

Butterfield, S. & Cutler, A. (1990). Intonational cues to word segmentation in clear speech? *Proceedings of the Institute of Acoustics,* 12, Part 10, 87-94.

Cutler, A. & Butterfield, S. (1990a). Durational cues to word boundaries in clear speech. *Speech Communication, 9,* 485-495.

Cutler, A. & Butterfield, S. (1990b). Syllabic lengthening as a word boundary cue. *Proceedings of the 3rd Australian International Conference on Speech Science and Technology,* pp.324-328.

Cutler, A. & Butterfield, S. (1991a). Rhythmic cues to speech segmentation: Evidence from juncture mispcrception. *Journal of Memory and Language,* 30, in press.

Cutler, A. & Butterfield, S. (1991b). Word boundary cues in clear speech: A supplementary report. *Speech Communication,* 10, in press.

Cutler, A. and Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2,*133-142.

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

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

Nooteboom, S.G. & Doodeman, G.J.N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America,* 67, 276-287.

Taft, L. (1984). *Prosodic Constraints and Lexical Parsing Strategies,* PhD Dissertation, University of Massachusetts.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America,* 58,434-445.