NATURAL SPEECH CUES TO WORD SEGMENTATION UNDER DIFFICULT LISTENING CONDITIONS

Anne Cutler and Sally Butterfield

MRC Applied Psychology Unit, 15 Chaucer Rd., Cambridge CB2 2EF, U.K

ABSTRACT

One of a listener's major tasks in understanding continuous speech is segmenting the speech signal into separate words. When listening conditions are difficult, speakers can help listeners by deliberately speaking more clearly. In three experiments, we examined how word boundaries are produced in deliberately clear speech. We found that speakers do indeed attempt to mark word boundaries; moreover, they differentiate between word boundaries in a way which suggests they are sensitive to listener needs. Application of heuristic segmentation strategies makes word boundaries before strong syllables easiest for listeners to perceive; but under difficult listening conditions speakers pay more attention to marking word boundaries before weak syllables, i.e. they mark those boundaries which are otherwise particularly hard to perceive.

INTRODUCTION

To understand continuous speech, a recogniser has to locate and identify parts of the speech signal which correspond to individual words. Unfortunately, segmenting continuous speech into words is not easy, since word boundaries are seldom explicitly marked. Human listeners respond to this problem by adopting various strategies to maximise the efficiency of word boundary location. For instance, word counts of spontaneously produced British English speech have shown that about 90% of lexical words (content words) begin with stressed (or, more exactly: strong) syllables [1]. Thus it would be a good bet to treat strong syllables as if they were highly likely to be word-initial. Indeed, listeners do segment English speech at the onset of strong syllables [2]. Moreover, when listeners misperceive word boundaries, their most likely mistake is the erroneous insertion of a boundary before a strong syllable [3].

However, human speakers can, if necessary, make word boundaries clear. For instance, a speaker could pause before every word. As we know, conversational speech is *never* like that. But speakers do speak in a range of styles, using careful articulation with foreigners, for example, but casual mumbles with close friends and family. And several recent studies have demonstrated that speakers who perceive that a listener is having difficulty do indeed adjust their speech towards clearer articulation when repeating. For instance, they speak more slowly, louder, and with raised pitch [4]; they simplify syntactic structure [5]; and they make segmental adjustments such as separating the VOT distributions for voiced and voiceless stop consonants and fully releasing stops in word-final position [6, 7].

Few such studies have examined precisely how word boundaries are produced when speakers arc deliberately trying to speak clearly (although there is evidence that clear speech contains pauses at word boundaries [7]). Studies of normal speech production have shown that speakers are reluctant to distort the initial boundaries of unpredictable or semantically focussed words, but happy to distort the initial boundaries of predictable words [8]; this suggests that not all word boundaries will necessarily be treated equally in clear speech.

In a series of experiments we have examined the question of whether speakers who know listening conditions are difficult try to make word boundaries clear. We further examined whether speakers distinguished between word boundaries preceding strong versus weak word-initial syllables, since the studies mentioned above have shown that this distinction is important to listeners. The present preliminary report is confined to two measures: pausing at a word boundary, and lengthening of the syllable preceding a boundary.

EXPERIMENT 1

We constructed 12 sentences of relatively unpredictable content. Each sentence contained a critical word boundary; in six sentences the word after this boundary began with a strong syllable, in six it began with a weak syllable. The sentences were paired so that phonetic material immediately either side of the boundary was comparable in a strong-syllable and a weak-syllable case. Examples are "Take it in turns to eat breakfast", where the critical boundary precedes "turns" (a strong syllable), versus "He called in to view it himself, where the critical boundary precedes "to" (here, a weak syllable).

For each sentence we also constructed two purported mishearings, to be presented to the subjects as "feedback".

These were chosen to be fairly realistic mishearings - for instance, the rhythm of the sentence was fairly well preserved, as were most of the vowels in the stressed syllables. In each case, however, the feedback sentences contained *no* boundary at the critical location. For the above examples, the feedback sentences were "Baker interns all the terrorists"/ "Take it internally at breakfast", and "The cold interviewer was selfish"/"He crawled into view by himself.

Five subjects took part (for payment) in the experiment. They were told that their speech was being fed through a distorting filter to a listener in the next room who would type what he thought he heard into a computer which in turn would display this response on the subjects' VDU screen. The subjects were asked to say each sentence as naturally as possible when first producing it. If the listener's response was incorrect, then the sentence should be repeated; if the response was still incorrect, it should be repeated yet again. Because for each experimental sentence the "listener's" response was indeed twice incorrect, mis instruction ensured that these sentences were produced three times each. The subjects were asked to speak clearly when repeating (but they were told not to shout as this would make the distortion worse).

Besides the 12 experimental sentences, subjects produced three practice and ten filler sentences, some of which the 'listener" apparently heard correctly on first or second hearing. All the subjects' productions were recorded.

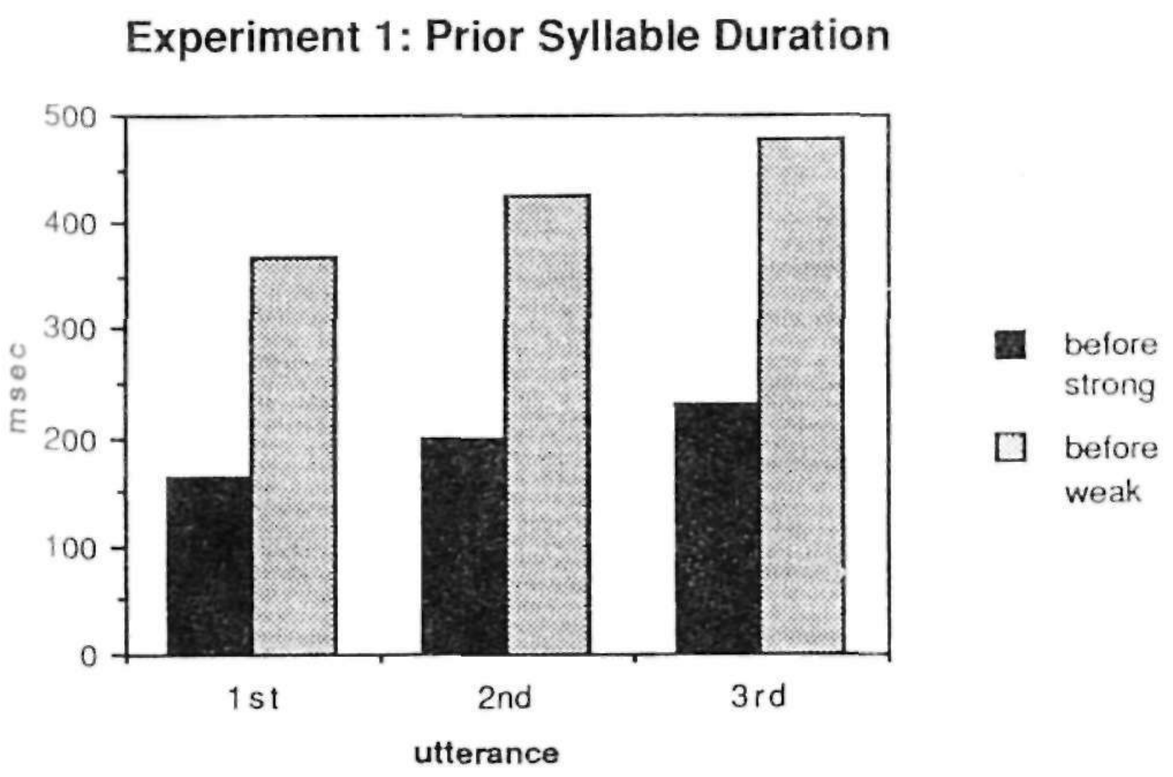
Each subject's three utterances of each of the 12 sentences were digitised - 180 utterances in all. This rich body of data lends itself to a variety of analyses. With respect solely to the critical word boundaries, we could examine whether there is a pause before the boundary; whether the syllable before the boundary is lengthened; whether the word-initial segments are articulated more clearly; whether word-final segments are differendy produced; and so on. These are time-consuming analyses, and many will be described in later reports; in this preliminary report we describe only two durational measures (of pauses and pre-boundary syllables).

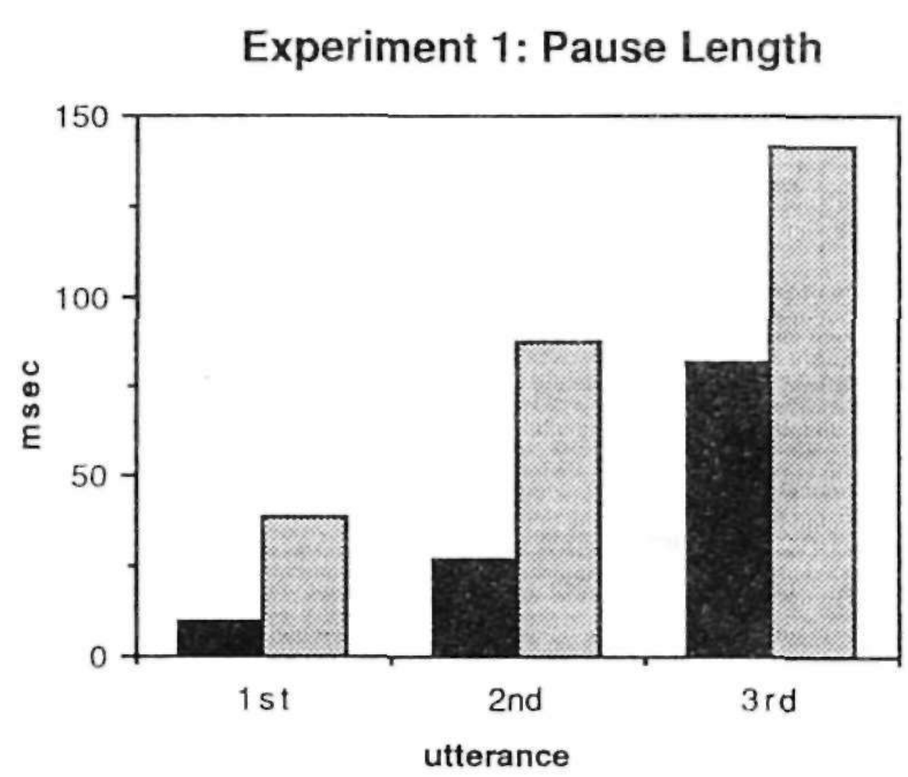
The figures for this experiment show the mean durations (across subjects and sentences) of the pauses and pre-boundary syllables in the first, second and third productions, separately for utterances where the boundary preceded a strong versus a weak syllable.

The greater length of pre-boundary syllables preceding weak word-initial syllables is an artefact of our materials; English has a tendency to alternate weak and strong syllables, and in some of our sentences weak syllables were preceded by strong syllables, and vice versa. Thus the proper measure to take is the amount of lengthening from first to second utterance, and from second to third. This showed that greater lengthening occurred before weak than before strong syllables (Fl [1,4] = 5.36, $p < .09$; F2 [1,10] = 5.05, $p < .05$).

Pauses also, are longer before weak syllables (Fl [1,4] = 2.63, $p > .1$, but F2 [1,10] = 5.91, $p < .04$).

A noteworthy feature of the results of this experiment is that there was no interaction between the effect of strong versus weak syllables and the effect of repetitions; as the graphs show, there was a difference between pre-strong and pre-weak boundaries even in the baseline utterances. Are these differences therefore characteristic of normal speech production, and not specific to deliberately clear speech? Since such effects have not previously been reported in studies of normal speech production, this suggestion seems unlikely. An alternative possibility is that subjects were speaking clearly even in the baseline utterances. In Experiment 2, therefore, we used as a baseline utterances which were collected before subjects were aware of the need to speak clearly.





EXPERIMENT 2

Five further subjects produced the same sentences under the same experimental conditions, with one exception: before being told about the listener and the supposed distortion, the subjects read the experimental and filler sentences aloud onto tape. These utterances served as the baseline, and were measured and compared with the two repetitions of each experimental sentence after feedback.

The figures for Experiment 2 again display the mean durations (across subjects and sentences) of the pauses and pre-boundary syllables in the baseline, second and third productions, separately for utterances where the boundary preceded a strong versus a weak syllable.

Again it can be seen that syllables preceding weak word-initial syllables are lengthened to a relatively greater degree than syllables preceding strong word-initial syllables, and the difference in increase is significant (Fl [1,4] = 12.54, $p < .03$; F2 [1,10] = 7.85, $p < .02$). However, the effect of strong versus weak syllables interacted significantly with the effect of repetitions, and subsequent t-tests showed that the increase from baseline to second utterance was significantly greater before a pre-weak than before a pre-strong boundary, but there was no significant difference between the two conditions in the increase from second to third utterance.

Pausing was again longer before a weak syllable (Fl [1,4] = 10.93, $p < .03$; F2 [1,10] = 2.56, $p > .1$), but the effects differed across repetitions: t-tests showed no difference in the baseline condition, but significantly longer pausing before weak than before strong syllables in repeated utterances.

Both Experiments 1 and 2 have suggested that speakers who are deliberately trying to speak clearly do indeed produce cues to the presence of a word boundary, and moreover, these cues tend to be more marked before a weak than before a strong initial syllable. However, it is possible that some differences between our sentence pairs might have contributed to the effects we found. For instance, as we pointed out above, the pre-boundary syllables were imperfectly matched. In addition, in some cases there was a difference between the two members of a pair in the syntactic strength of the crucial boundary, although the differences involved minor phrases boundaries, which are not usually marked in normal speech production [8]. In Experiment 3 we attempted to control for these possible confounding effects.

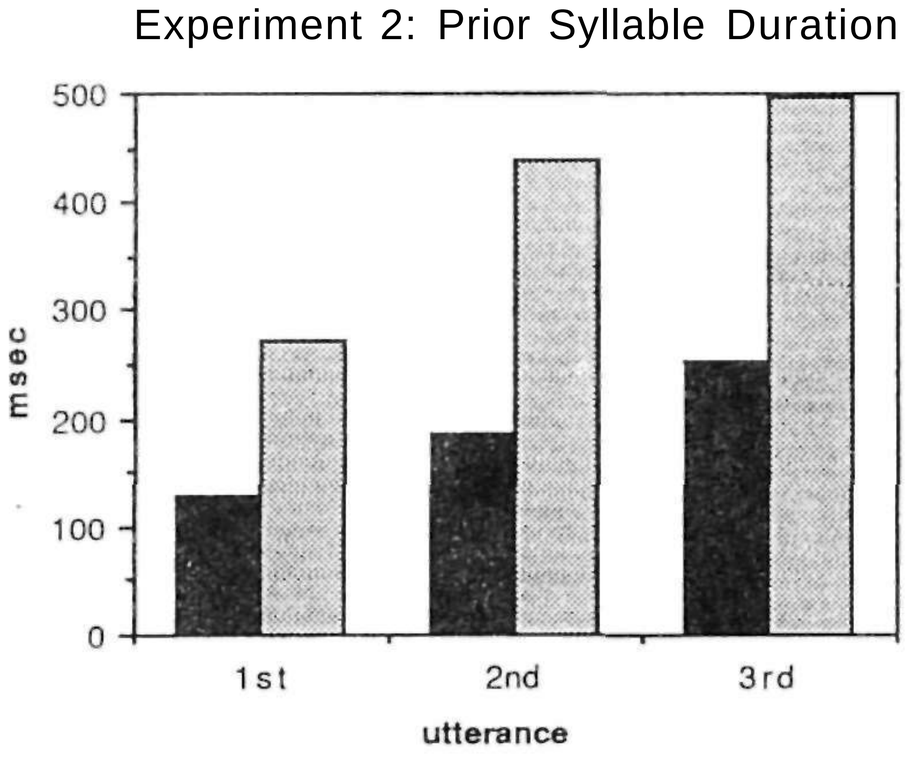
EXPERIMENT 3

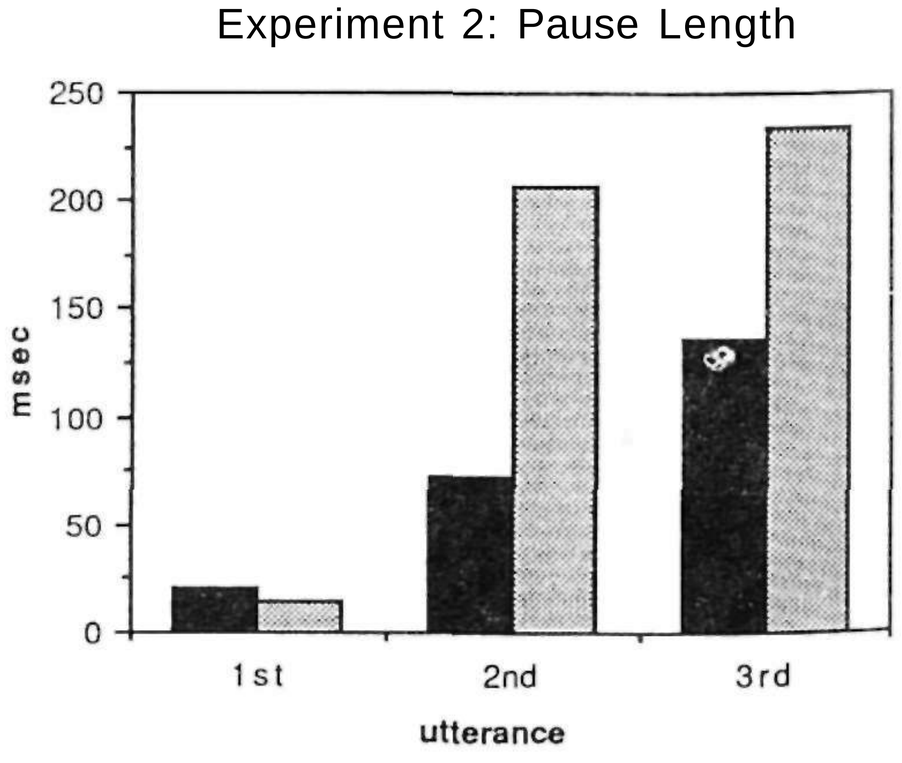
A further set of 12 sentences was constructed, again in 6 matched pairs containing strong and weak syllables after the critical boundary. Word class of the word after the boundary was matched in each pair, as was syntactic strength of the boundary and identity of the pre-boundary syllable. An example pair is "Play this card a good deal more "/"Fire this cadet's automatic"; the crucial boundary is "this c-". Purported mishearings were again constructed for use as feedback.

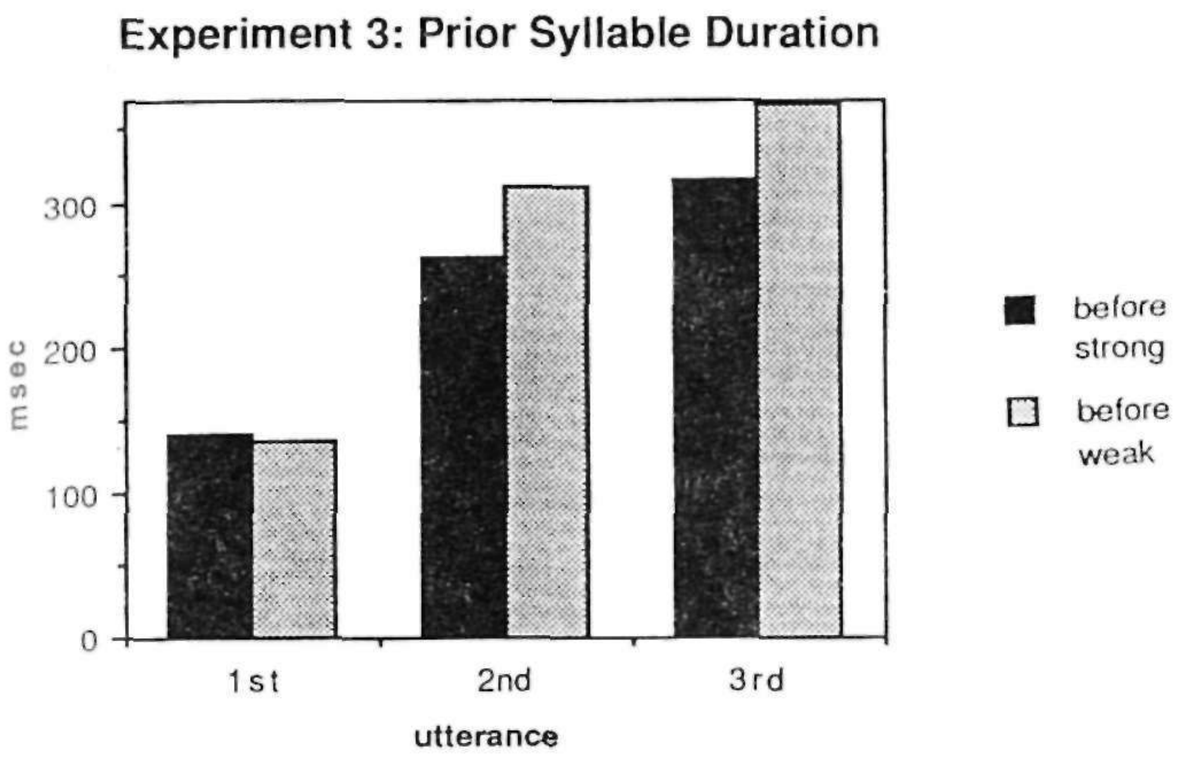
Five subjects took part; the conditions were as in Experiment 2.

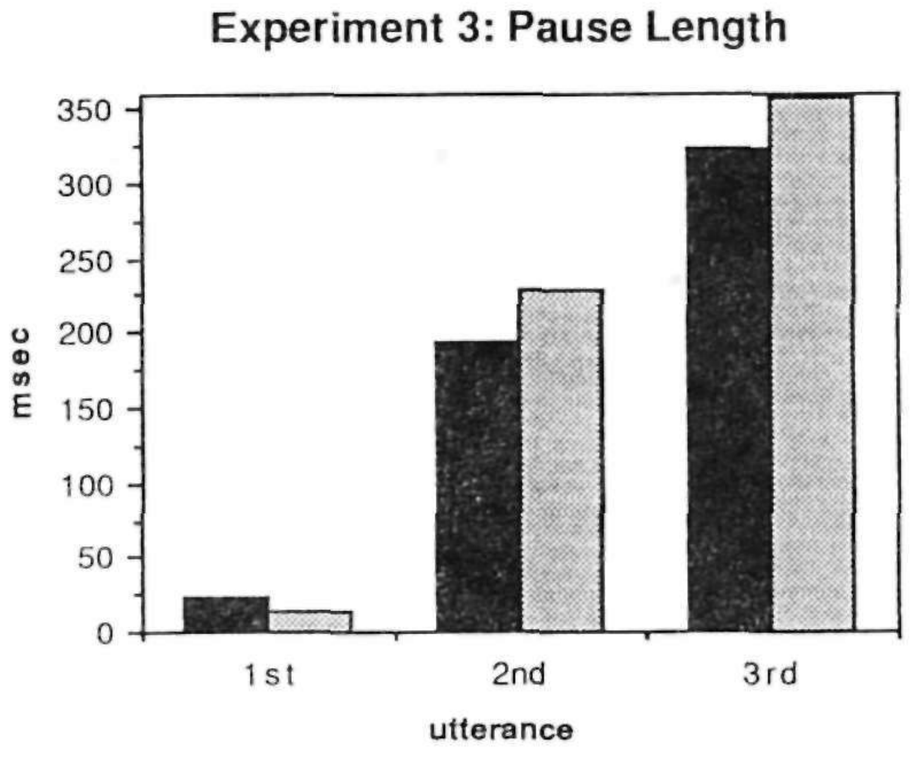
The figures for Experiment 3 again show the mean durations (across subjects and sentences) of the pauses and pre-boundary syllables in the baseline, second and third productions, separately for utterances where the boundary preceded a strong versus a weak syllable.

The results were very similar to those of Experiment 2. The duration of the pre-boundary syllable (recall that in this experiment the pre-boundary syllable is the same in each member of a pair) hardly differs in the baseline, but is lengthened by a much greater amount before weak









syllables than before strong in the repetitions; t-tests again show that the increase from baseline to second utterance is significantly greater before weak than before strong syllables, but there is no significant difference between the conditions in the amount of increase from second to third utterance.

Likewise, the pauses show no difference in the baseline utterance, but in the repetitions pauses preceding weak syllables are longer than pauses preceding strong syllables (although in this case the effect fails to reach statistical significance).

CONCLUSION

Firstly, we have shown that speakers do mark word boundaries when they are trying to produce clear speech for a listener's benefit. The best evidence for this comes from the analysis of pausing in Experiments 1 and 2. In the baseline utterances, when speakers had no knowledge of a need to speak clearly, they virtually never paused at a word boundary, but in their repetitions they paused before words for an average of 150 to 250 ms. This suggests that speakers are aware that segmentation of continuous speech into words is a major problem for the listener.

Secondly, we have shown that speakers differentiate between boundaries which precede strong and weak syllables: on both the measures we examined, boundaries preceding weak syllables were in general marked more clearly than boundaries preceding strong syllables. Given that perceptual evidence has shown that listeners (to English) adopt a strategy of segmenting speech at strong syllable onsets, mis result suggests that speakers may compensate for listener strategies by enhancing the clarity particularly of those word boundaries which would *not* be identified by application of the customary strategies. In both cases, therefore, our results suggest that speakers are able to tailor their speech very efficiently to their listeners' needs.

ACKNOWLEDGEMENT

This research was supported by IBM UK Scientific Centre.

REFERENCES

[1] Cutler, A. & Carter, D. (1987) The prosodic structure of initial syllables in English. *Proceedings of the European Conference on Speech Technology,* Edinburgh; Vol. 1, pp. 207-210.

[2] 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.

[3] Butterfield, S. and Cutler, A. (1988) Segmentation errors by human listeners: Evidence for a prosodic segmentation strategy. *Proceedings of SPEECH '88* (Seventh Conference of the Federation of Acoustic Societies of Europe), Edinburgh; Vol. 3, pp. 827-833.

[4] Clark, J., Lubker, J. & Hunnicutt, S. (1988) Some preliminary evidence for phonetic adjustment strategies in communication difficulty. In R. Steele & T. Threadgold (Eds.) *Language Topics: Essays in Honour of Michael Halliday.* Amsterdam: John Benjamins; pp. 161-180.

[5] Valian, V.V. & Wales, R.J. (1976) What's what: talkers help listeners hear and understand by clarifying syntactic reladons. *Cognition, 4,* 115-176.

[6] Chen, F.R., Zue, V.W., Picheny, M.A., Durlach, N.I. & Braida, L.D. (1983) Speaking clearly: Acoustic characteristics and intelligibility of stop consonants. *Speech Communication Group, MIT: Working Papers, 2,*1-8.

[7] Picheny, M.A., Durlach, N.I. & Braida, L.D. (1986) Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research, 29,* 434-446.

[8] Cooper, W.E. & Paccia-Cooper, J. (1980) *Syntax and Soeech.* Cambridge. MA: Harvard Univ. Press.