Proceedings of the Institute of Acoustics

INTONATIONAL CUES TO WORD BOUNDARIES IN CLEAR SPEECH?

Sally Butterfield & Anne Cutler

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

1. INTRODUCTION

To understand continuous speech, listeners have to locate and identify parts of the speech signal which correspond to individual words. However, segmenting continuous speech into words is a non-trivial task, because robust and obligatory cues to the presence of word boundaries are not a feature of natural speech.

However, human speakers can, if necessary, adjust their speaking style, using careful articulation with foreigners, for example, but casual mumbles with close friends and family. And several recent studies have demonstrated that speakers who notice that a listener is having difficulty do indeed adjust their speech towards clearer articulation when repeating. Adjustments include speaking more slowly, louder, and with raised pitch [1]; making syntactic structure explicit [2]; and implementing segmental changes such as separating the VOT distributions for voiced and voiceless stop consonants and fully releasing word-final stops [3, 4].

In our laboratory we have examined precisely how word boundaries are produced when speakers are deliberately trying to speak clearly. When speakers know that listening conditions are difficult, they may pay particular attention to helping listeners with the segmentation problem, by trying hard to make word boundaries clear; moreover, they may distinguish between types of word boundaries by making some even clearer than others. Picheny, Durlach and Braida [4] found that clear speech contains pauses at word boundaries, although most such pauses were much shorter than the 250 ms which is commonly used as the threshold for defining a pause in other studies [5]. They did not, in their report, distinguish between *types* of word boundaries. Studies of normal speech production, however, suggest that not all word boundaries will necessarily be treated equally. Cooper and Paccia-Cooper [6] studied the application across word boundaries of phonological rules such as palatalisation, and found that speakers are reluctant to apply such rules when they will distort the initial boundaries of low frequency or contrastively stressed words; however, they are happy to apply them across the initial boundaries of high frequency, unstressed words.

In normal speech recognition, too, listeners differentiate between types of word boundaries. In English, more words begin with strong syllables (in which the vowel quality of the nucleus is full) than with weak syllables (in which the vowel quality of the nucleus is reduced; [7]). Knowledge of this bias in vocabulary structure apparently guides human listeners' strategies for dealing with the problem of word boundary location. For instance, listeners segment English nonsense strings at the onset of strong syllables [8], and when listeners misperceive word boundaries, their most likely mistake is the erroneous insertion of a boundary before a strong syllable [9].

If the distinction between strong and weak word-initial syllables is important for speech segmentation, it is reasonable to ask whether this distinction is also relevant in describing clear speech phenomena. In the present research programme we have investigated whether a distinction is made in clear speech between boundaries preceding strong versus weak syllables.

We reported previously [10, 11] that durational cues (pause insertion and pre-boundary lengthening) were greater for boundaries preceding weak syllables (e.g "in / to") than for boundaries preceding strong syllables (e.g. "in / turns"). In English, listeners tend to segment speech at the onset of strong syllables; we argued, therefore, that speakers' clear-speech adjustments mark just those boundaries which would *not* be perceived by application of this usual procedure. However, enhanced durational boundary cues before weak syllables may merely compensate for the lack of alternative means of marking such a boundary; the greater possibility of intonational variation on strong syllables may allow word-initial strong syllables to be sufficiently clearly signalled by enhanced pitch movement in clear speech that no other boundary marking is required. The present study tests this hypothesis.

2. EXPERIMENTS 1 and 2

2.1 Method

2.1.1 Materials. Twelve sentences of relatively unpredictable content were constructed. 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).

The form of the sentences was in part determined by the availability of possible mishearings in which the critical boundary was absent. For instance, "in turns" could be misheard as "interns", while "in to" could be misheard as "into". For each sentence we constructed two such purported mishearings, to be presented to the subjects as feedback. These were quite realistic as mishearings - 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" and "Take it internally at breakfast", versus "The cold interviewer was selfish" and "He crawled into view by himself". The full set of target and feedback sentences may be found in Cutler and Butterfield [11].

2.1.2 Subjects and Procedure. In Experiment 1, five members of the Applied Psychology Unit subject panel 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; this response would be displayed on the subjects' VDU. In fact the only listener was the experimenter, and all subjects received the same "feedback". The subjects were given the sentences on cards, and were asked to read 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 second response was again incorrect, the sentence should be repeated once more. Because for each experimental sentence the "listener's" response was indeed twice incorrect, this 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.

In Experiment 2, five further subjects from the same population produced the same sentences under the same conditions, with one exception: the subjects read the experimental (and filler) sentences aloud onto tape before they were told about the listener and the supposed distortion. These initial productions then served as the baseline to be compared with the two post-feedback repetitions.

For each subject in each experiment, the baseline production and both repetitions of each experimental sentence were digitised at a sampling rate of 10 kHz. The syllable before and after each critical boundary was marked and measured. Previously [10, 11] we reported two durational measures (of pauses and of pre-boundary syllables). The present report describes analysis of the pitch contours of the syllables before and after the boundary. Each syllable was analyzed by the Schafer-Vincent algorithm [12]; this algorithm calculates a fundamental frequency (F0) by detecting quasi-periodic parts of a speech signal and analyzing the structure of an amplitude-against-time representation of the signal. The F0 values within each syllable were averaged, and a standard deviation calculated.

The algorithm failed to calculate values for 10.8% of the syllables. In these cases the missing data point was replaced by the subject's average for that condition.

2.2 Results and Discussion

The mean F0 value across each syllable gives an estimate of relative pitch *height*; the standard deviation of this mean for each syllable gives an estimate of relative pitch *movement* within the syllable. A separate analysis of variance was carried out on each measure.

The analysis of the mean F0 values revealed no significant differences of interest in either experiment. In neither experiment was there a significant difference in mean F0 on either the syllable before or the syllable after the boundary as a function of whether the boundary preceded a strong versus weak syllable. The only significant effects were a tendency for mean F0 of syllables before the boundary to become *lower* across repetitions (F1 [2,8] = 5.45, p < .04) in Experiment 1 only, and a tendency for syllables after the boundary to have lower F0 than syllables before the boundary (F1 [1,4] = 16.77, p < .02), again in Experiment 1 only. (This latter effect is presumably due to pitch declination.)

The analysis of standard deviations in Experiment 1 showed that there was more movement on pre-boundary syllables if the boundary preceded a weak syllable (F1 [1,4] = 15.81, p < .02), but more movement on post-boundary syllables when these syllables were strong (F1 [1,4] = 21.5, p < .01). Since pre-boundary syllables were usually themselves strong when the boundary preceded a weak syllable, these results simply show that, as expected, there is more pitch movement on strong syllables. The effect did not interact with the repetitions factor, and *t*-tests showed that in each case it was significant even in baseline utterances (*t* [4] = -3.05, p < .04 for pre-boundary syllables, *t* [4] = 4.25, p < .02 for post-boundary syllables). The same two main effects also showed up in Experiment 2 (F1 [1,4] = 11.52, p < .03, and F1 [1,4] = 21.41, p < .01, respectively). In this experiment both effects were, however, significantly stronger in repeated than in baseline utterances, although again *t*-tests showed that in baseline utterances the effects nevertheless approached significance (*t* [4] = -2.63, p < .06 for pre-boundary syllables, *t* [4] = 2.24, p < .09 for post-boundary syllables).

The implications of these results are difficult to determine. More pitch movement occurs on strong syllables than on weak, and, at least in Experiment 2, this difference is even more marked in clear than in baseline utterances. But there is little suggestion that the pitch movement is of particular use to lexical segmentation - when the post-boundary syllable is strong, it carries more pitch movement, but when the post-boundary syllable is weak, it is preceded by more pitch movement.

The difficulty of interpretation is partly a function of the fact that the speech materials used in Experiments 1 and 2 were not necessarily matched on all relevant dimensions. For instance, there may have been an effect of word frequency, such as Cooper and Paccia-Cooper [6] found for natural speech. When the members of each pair were compared on word frequency using the Francis and Kučera [13] word-class-specific norms, frequency was found to be quite strongly mismatched, because of a word class mismatch: five of the six weak syllables were high frequency closed class words. Thus the frequency of occurrence for the weak post-boundary syllables was in general much higher than that of the strong syllables. It may be the case that low frequency words, irrespective of whether they are realised as strong or weak syllables, attract pitch movement. Alternatively, it may be the case that closed class words are not seen as a suitable domain for pitch movement whereas open class words are. Closer control of both word class and frequency is therefore necessary to provide more readily interpretable data.

Accordingly, two further experiments were conducted. In Experiment 3, we manipulated prosodic structure while keeping word class constant, and in Experiment 4 we manipulated word class - the closed/open distinction - while keeping prosodic structure as far as possible constant. Because closed class words are always of higher average frequency than open class words, the frequency difference in Experiment 4 was in the same direction as in Experiments 1 and 2, and therefore in Experiment 3 we manipulated frequency in the opposite direction - all initially-strong words were of higher frequency than their initially-weak pairs. In particular, note that Experiment 3 offers a more closely controlled investigation of the strong-weak comparison addressed in Experiments 1 and 2.

3. EXPERIMENT 3

3.1 Method

3.1.1 Materials. A further set of 12 sentences was constructed, again in six 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-". All the words with strong initial syllables were higher in frequency of occurrence than their weak-initial pairs. Purported mishearings were again constructed for use as feedback. The complete set of target and feedback sentences is listed in Cutler and Butterfield [11].

3.1.2 Subjects and Procedure. Ten subjects from the same population took part; the procedure was as in Experiment 2. This experiment and the next were administered together, so that the total number of items, including the three practice and ten filler sentences, was 37.

3.2 Results and Discussion

Algorithm failures (18.6% of cases) were replaced in the same manner as for the preceding experiments. The analysis of mean F0 revealed no significant effects at all for pre-boundary syllables (which were matched across the strong-weak comparison in this experiment). For post-boundary syllables there was a tendency for higher F0 to be used in clear than in baseline utterances, and this interacted with the repetitions variable: *t*-tests showed the source of this interaction to be a significantly higher mean F0 on weak than on strong syllables in baseline productions (t [9] = -2.93, p < .02) but no difference in clear utterances (t < 1 in both cases).

The analysis of the standard deviations on the syllable preceding the boundary again showed no differences as a function of boundary type; there was more movement in clear than in baseline utterances, but no effect either in baseline or clear utterances of whether the boundary preceded a strong or weak syllable. On the post-boundary syllables, which of course did differ, there was more movement on strong than on weak syllables (F [1,9] = 29.93, p < .001), but this effect did not interact with the repetitions variable, and *t*-tests showed that the difference was significant in baseline (*t* [9] = 2.84, p < .02) as well as in clear utterances (*t* [9] = 6.97 and 4.98, both p < .001).

Thus this experiment suggests that in general, more pitch movement occurs on strong than on weak syllables, but this is true of any utterance, whether or not the speaker is aiming at clear articulation - in other words, cues to lexical segmentation in clear speech do not exploit F0. Word frequency does not appear to play a role, since the difference between strong and weak syllables was similar in Experiments 1, 2 and 3, although the frequency difference between strong and weak syllables in Experiment 3 was the reverse of that in Experiments 1 and 2. Experiment 4 further investigates the possible contribution of word class to the results of Experiments 1 and 2.

4. EXPERIMENT 4

4.1 Method

4.1.1 Materials. A further set of 12 sentences was constructed, again in six matched pairs. In this case the critical variable was word class of the post-boundary word; homophones were chosen which could be either open- or closed-class words. An example pair is *hour/our*, as in "Lots of hour-long sessions are needed" versus "Both of our children like peanuts"; the crucial boundary is "of (h)our". Each closed class word was higher in frequency than its open class pair. Although it would have been desirable to vary word class fully independently of the strong/weak syllable distinction, this is impossible because nearly all closed class words, but no open class words, can be reduced in sentence contexts; most closed class words which cannot be reduced - *these, those* etc. - are not homophonous with open class words. Where we could, we chose homophones which could not be reduced, and for the remaining items we chose contexts in which reduction was unlikely.

Purported mishearings were again constructed for use as feedback. The complete set of target and feedback sentences is listed in Cutler and Butterfield [11].

4.1.2 Subjects and Procedure. This experiment was administered together with Experiment 3.

4.2 Results and Discussion

Algorithm failures (11% of cases) were replaced in the same manner as for the preceding experiments. The analysis of mean F0 showed higher F0 when the homophone was an open- rather than a closed-class word, on both pre- and post-boundary syllables (F [1,9] = 41.52 and 95.22 respectively, both p < .001); in neither case did this effect interact with the repetitions variable.

The analysis of standard deviations showed no significant effects at all for pre-boundary syllables. Post-boundary syllables showed more movement on open- than on closed-class words (F [1,9] = 29.15, p < .001), and more movement in clear than in baseline utterances (F [2,18] = 21.64, p < .001), but again these two effects did not interact, and *t*-tests showed that the open-closed difference was significant in baseline (*t* [9] = -3.29, p < .01) as well as in clear utterances (*t* [9] = -4.01 and -3.95, both p < .01).

This experiment certainly suggests that pitch movement is more likely if a given syllable is functioning as an open-class rather than a closed-class word. Word class effects may therefore have played some role in Experiments 1 and 2. The most important result of the present study, however, is that we again find that the effects which obtain in clear speech are also present in the baseline utterances. Thus this experiment provides further evidence that F0 does not serve as a cue to lexical segmentation in clear speech.

5. CONCLUSION

The results reported here should be interpreted in conjunction with our durational analyses of clear speech [10, 11]. In those analyses we found strong effects of the nature of the word boundary which speakers were attempting to make clear: durational signals (pausing and lengthening) were significantly more marked for boundaries preceding weak than for boundaries preceding strong syllables.

As we argued in the introduction, it is conceivable that there could be a trade-off between different sources of information such that some types of lexical boundary are more readily signalled in one way, other boundary types in some alternative way. Intonational variation is an obvious candidate for an alternative source of information in the present instance, since there is more opportunity for pitch movement on strong syllables, and hence more opportunity for a speaker who is deliberately trying to speak clearly to exploit intonation to signal boundaries preceding strong syllables.

The results, however, suggest that if there is such a trade-off, it does not include intonation. Our analysis showed that more pitch movement occurred on strong than on weak syllables, as one would expect, but this pattern was seen both in baseline and in clear utterances. There were increases in pitch movement in clear speech, but these were not significantly greater for strong versus weak word-initial syllables. We conclude that word boundary cues in clear speech exploit duration but not intonation. Where a source of information can be exploited to signal a word boundary - i.e. in the durational case - the evidence suggests that speakers consider boundaries before weak syllables more in need of marking than boundaries before strong syllables.

6. ACKNOWLEDGEMENTS

This research was supported by IBM UK Scientific Centre. We are grateful for assistance from Brian Pickering, John Holdsworth and Brit van Ooyen.

7. REFERENCES

[1] J CLARK, J LUBKER & S HUNNICUTT, 'Some preliminary evidence for phonetic adjustment strategies in communication difficulty', in *Language Topics: Essays in Honour of Michael Halliday*, ed. R. Steele & T. Threadgold. Amsterdam: John Benjamins, p161 (1988)

[2] V V VALIAN & R J WALES, 'What's what: talkers help listeners hear and understand by clarifying syntactic relations', *Cognition*, 4, p115 (1976)

[3] F R CHEN, V W ZUE, M A PICHENY, N I DURLACH & L D BRAIDA, 'Speaking clearly: Acoustic characteristics and intelligibility of stop consonants', *Speech Communication Group, MIT: Working Papers*, 2, p1 (1983)

[4] M A PICHENY, N I DURLACH & L D BRAIDA, 'Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech', *J Sp & Hear Res*, 29, p434 (1986)

[5] F GROSJEAN, 'Linguistic structures and performance structures: Studies in pause distribution', in *Temporal Variables in Speech*, ed. H.W. Dechert & M. Raupach. The Hague: Mouton, p91 (1980)

[6] W E COOPER & J M PACCIA-COOPER, *Syntax and Speech*. Cambridge, MA: Harvard Univ. Press (1980)

[7] A CUTLER & D CARTER, 'The predominance of strong initial syllables in the English vocabulary', *Comp Sp & Lang*, 2, p133 (1987)

[8] A CUTLER & D G NORRIS, 'The role of strong syllables in segmentation for lexical access', *J Exp Psy: Hum Perc & Perf*, 14, p113 (1988)

[9] S BUTTERFIELD & A CUTLER, 'Segmentation errors by human listeners: Evidence for a prosodic segmentation strategy', *Proc. SPEECH '88*, Vol. 3, p827 (1988)

[10] A CUTLER & S BUTTERFIELD, 'Natural speech cues to word segmentation under difficult listening conditions', *Proc. EUROSPEECH '89*, Vol. 2, p.372 (1989)

[11] A CUTLER & S BUTTERFIELD, 'Durational cues to word boundaries in clear speech', *Speech Comm*, 9 (1990)

[12] K SCHAFER-VINCENT, 'Pitch period detection and chaining: Method and evaluation', *Phonetica*, 40, p177 (1983)

[13] W N FRANCIS & H KUČERA, *Frequency Analysis of English Usage*. Boston: Houghton Mifflin (1982)