PROCEEDINGS OF THE THIRD AUSTRALIAN INTERNATIONAL CONFERENCE ON

SPEECH SCIENCE AND TECHNOLOGY

Melbourne
27-29 November 1990

Edited by

Roland Seidl
Telecom Research Labs

Australian Speech Science and Technology Association
Canberra 1990

SYLLABIC LENGTHENING AS A WORD BOUNDARY CUE

Anne Cutler and Sally Butterfield

MRC Applied Psychology Unit, Cambridge, UK.

ABSTRACT - Bisyllabic sequences which could be interpreted as one word or two were produced in sentence contexts by a trained speaker, and syllabic durations measured. Listeners judged whether the bisyllables, excised from context, were one word or two. The proportion of two-word choices correlated positively with measured duration, but only for bisyllables stressed on the second syllable. The results may suggest a limit for listener sensitivity to syllabic lengthening as a word boundary cue.

INTRODUCTION

Our memories do not have infinite capacity; we could not possibly store a representation of every utterance we might ever hear. Therefore, to *recognise* a spoken utterance, we must recognise those of its individual components which do correspond to items which we have stored in memory. To do that, we must also work out how the incoming speech stream divides up into separate components, i.e. we must locate word boundaries. Unfortunately, segmenting continuous speech into words is not easy, since there appear to be no robust obligatory cues to the presence of a boundary at the word level.

In previous studies in our laboratory we have investigated how speakers help listeners by speaking more clearly when listening conditions are obviously difficult. We found (Butterfield & Cutler, 1990; Cutler & Butterfield, 1989, 1990) that speakers do attempt to mark word boundaries, but that they do not use any and every means of doing so which they potentially could. Specifically, intonational cues are not exploited, but durational cues are. Speakers exploit several durational manipulations to make word boundaries clearer, including increase in the duration of the syllable before the boundary, and insertion of a pause at the boundary.

If such manipulations are to be of use to listeners, however, the capacity to exploit variation of this nature must 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. It is unclear, however, whether pre-boundary syllable lengthening exists outside of deliberately clear speech, and if so, whether listeners exploit it to perform lexical segmentation. The present study attempts to address each of these issues.

The evidence from previous studies is mixed. The classic study of Lehiste (1972) found no difference in the first syllables of bisyllabic strings such as *speeder* and *speed kills;* Lehiste 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. Perhaps because of these negative results, most recent studies of segment and syllable durations do not contrast word-final versus non-word-final syllables. Beckman and Edwards (1990), however, report word-final lengthening in juncturally ambiguous strings such as "Pop opposed" versus "Poppa posed".

In typical English speech, most words are monosyllabic (Cutler & Carter, 1987), which means that there would be very little opportunity for speakers to differentiate between word-final and non-word-final syllables, and likewise very little opportunity for listeners to exploit such differentiation. Perhaps ambiguous strings of the kind used by Beckman and Edwards offer the greatest chance of observing word-final lengthening, both in production and perception. Such ambiguous sequences were indeed 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. Each item was spoken only once, and Taft made no measurements of her productions. The subjects showed a general bias towards one-word choices, but items with final stress (e.g. *invests)* received more two-word choices than items with initial stress.

Taft's experiment does not directly address the issues which concern the present study, but it offers a methodological precedent. It also suggests that there may be listener biases which operate independently of the acoustic information in the signal. Taft interpreted her results as evidence that, in English, listeners prefer word boundaries to precede stressed syllables, a claim for which there is more recent experimental support (Cutler & Norris, 1988; Butterfield & Cutler, 1988). In the present study 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.

PRODUCTION DATA

Method

24 ambiguous bisyllabic strings were chosen, each of which could be one or two words; twelve items had initial stress (e.g. *lettuce/let us),* twelve had final stress (e.g. *inquires/in choirs).* The items are listed in Table 1. All items were embedded in final position in sentence contexts (e.g. "We bought some apples and a beautiful lettuce; we'll go home early if our duties will let us"), and recorded by a phonetically trained speaker of British English. Each sentence was recorded twice: once spoken in the way the speaker deemed most natural for that sentence, and once with the prosodic contour of the sentence matched to the other sentence in that pair. These renditions were termed the "natural" and "unnatural" productions respectively. Thus for "We bought some apples and a beautiful lettuce" the "unnatural" production had the prosodic contour which the speaker had used for "We'll go home early if our duties will let us".

The bisyllables were excised from context and digitised at a sampling rate of 10 kHz, and the duration of both syllables in each token was measured on the digitised waveform.

Results and Discussion

The mean durations (averaged across natural and unnatural productions) for each syllable of each item are given in Table 1. An analysis of variance was carried out on the measurements for first and second syllables separately. The mean length of the first syllable of final-stress items was 78.5 ms when they were spoken as one word, 84 ms when they were spoken as two; for initial-stress items the means were 245 and 250 ms respectively. The difference in duration between one- and two-word productions was significant ($F [1,22] = 5.7$, $p < .03$), but this effect did not interact with stress pattern. The mean length of the second syllable of final-stress items was 448 ms when they were spoken as one word, 458 ms when they were spoken as two; for initial-stress items the means were 184 and 183 ms respectively. The difference in duration between one- and two-word productions was not significant overall, but the interaction of this effect with stress pattern was significant ($F [1,22] = 4.48$, $p<.05$).

Thus in final-stress items both syllables were lengthened in the two-word version compared to the one-word, but in initial-stress items boundary-conditioned lengthening was confined to the first syllable.

The contrast between "natural" and "unnatural" productions was not significant and did not interact with any other variable. The differences between one- and two-word productions were reduced in the "unnatural" productions, but were in the same direction as in the "natural". This suggests that the speaker was unsuccessful in matching his productions; had he succeeded, the unnatural two-word productions should have sounded like the natural one-word, for instance. Since this contrast had no significant effects, it was not explored further.

A correlation analysis was carried out on the two syllable measurements across items, to determine whether there were compensatory effects operating to keep overall duration of the bisyllable roughly constant. If so, lengthening of one syllable should accompany shortening of the other, and the

correlation would be negative. For initial-stress items the correlation was indeed negative (r [47] = -.15), but insignificantly so. For final-stress items, however, the correlation was significant and *positive* (r [47] = .56, p < .001); that is, longer first syllables tended to accompany longer second syllables. Within each stress pattern, the effects did not differ for one- versus two-word productions.

Since compensatory durational effects are usually explained as occurring at the foot level (Beckman & Edwards, 1990), the difference between item types is not surprising: in initial-stress items both syllables are in the same foot, but in final-stress items a foot boundary occurs prior to the second, stressed, syllable, and hence the two syllables belong to different feet. Thus compensatory effects, if they are present in such utterances, should be expected only in the initial-stress case.

PERCEPTION DATA

Method

A tape was made containing all the ambiguous bisyllables which had been excised from the sentence contexts. All four versions of each bisyllable (one-word and two-word versions, in natural and unnatural renditions) occurred on the tape; order of occurrence of the versions was counterbalanced across four subsets of the items also balanced for stress pattern. Response sheets were constructed which contained the two full sentence contexts for each bisyllable on the tape.

Ambiguous Bisyllables					
One-word Version			Two-word Version		
	First Syllable Duration	Second Syllable Duration		First Syllable Duration	Second Syllable Duration
Initial-stress					
lettuce	210	271	let us	221	262
office	153	259	off us	153	249
service	315	273	serve us	315	291
budget	202	192	budge it	198	210
market	312	178	mark it	317	181
packet	195	167	pack it	205	151
border	216	133	bored 'er	208	141
fitter	210	135	fit 'er	233	130
lever	281	131	leave 'er	274	132
forum	309	162	for 'em	330	147
freedom	275	159	freed 'em	274	132
tandem	264	150	tanned 'em	274	151
Final-stress					
about	29	403	a bout	27	404
assign	38	436	a sign	43	453
attest	37	490	a test	39	479
before	69	359	be four	73	377
beheaded	66	340	be headed	72	359
below	44	323	be low	62	319
infers	141	529	in furs	169	517
inquires	127	590	in choirs	147	622
invests	148	558	in vests	153	567
terrain	95	348	to rain	88	365
today's	66	497	to days	56	508
towards	86	505	to wards	83	530

Table 1. Ambiguous bisyllables used in the study, with measured durations (in ms) for each syllable in both one- and two-word versions produced by the speaker.

Subjects were 32 members of the Applied Psychology Unit subject panel, who were paid for participating. 30 subjects were native speakers of British English and only the responses of these 30 were analysed. The subjects were tested in two groups of 16. The subjects listened to the tape and, for each bisyllable which they heard (in isolation), they judged which context they thought it had come from (i.e., whether they judged it to be the one- or two-word version).

Results and Discussion

As in Taft's experiment, the listeners showed a general preference for making one-word choices; 67% of choices were for the one-word version. This presumably reflects properties of the task; an item presented in isolation is more likely to be expected to be one word than two. An analysis of variance was carried out on the number of two-word choices received by each item (note that it is an arbitrary decision to analyse two-word choices as opposed to one-word; the two alternatives are mirror images and either analysis would give the same result). Only one effect reached significance: subjects made significantly more two-word choices to two-word productions (36.4%) than to one-word productions (29.5%; $F [1,22] = 9.37$, $p < .01$). Neither stress pattern nor naturalness had any effect either alone or in interaction. Thus subjects were to some extent able to distinguish the speaker's one- versus two-word productions in spite of the tendency of the task to elicit a bias towards one-word choices. Note, however, that we did not replicate Taft's finding that final-stress items tended to elicit more two-word choices.

Our final analysis correlated the proportion of two-word choices received by each item with the item's measured duration. For final-stress items (e.g. *inquires/in choirs),* the proportion of two-word choices correlated positively with measured duration, both for first syllable durations ($r [47] = .46$, $p < .001$) and for second syllable durations ($r [47] = .55$, $p < .001$). (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 initial-stress items (e.g. *lettuce/let us),* however, no correlations with any durational measure were statistically significant.

The positive correlations indicate that greater item duration was associated with a higher likelihood that listeners would choose the two-word response alternative; however, this association was present only in the case of bisyllables with final stress.

CONCLUSION

This study was motivated by the observation that speakers producing deliberately clear speech appear to lengthen pre-boundary syllables as a way of marking word boundaries. Although clear speech is overall much slower than normal speech, the relative lengthening is greater on syllables preceding word boundaries which speakers are particularly trying to mark (as evidenced, for example, by greater pausing at those boundaries). We set out to examine whether such boundary-conditioned lengthening phenomena exist in normal as well as in clear speech, and whether listeners can exploit such information in making decisions about speech segmentation. Because spoken English typically contains more monosyllables than polysyllables, and hence does not offer much scope for contrasting word-final with non-word-final syllables, we investigated the production and perception of juncturally ambiguous bisyllables, i.e. two syllables which could form one word or two.

Our study has produced evidence that spoken productions of such bisyllables do indeed show some temporal differences, and that listeners can make use of these 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. Given that normal English speech offers so little scope for boundary-conditioned lengthening, it may play little role in normal speech recognition. A further reason why it may be of limited use is that any temporal cue can only be interpreted relative to the temporal pattern of the utterance in which it occurs. Thus it is not surprising to find that our subjects' accuracy of identification improved from the first to the second half of the experiment, presumably as they adjusted to the rate of speech used by the single speaker.

The improvement, however, was confined to final-stress items; initial-stress items showed no improvement. This is perhaps unsurprising since the perception data showed that it was only for the final-stress items that subjects' responses were determined by duration. Of course, the production data also showed that boundary-conditioned lengthening was larger in the case of final-stress items. We suggest that only in the final-stress case did the boundary-conditioned lengthening reach listeners' thresholds for durational discrimination. Nooteboom and Doodeman (1980) found that the just noticeable difference for vowel duration discriminating the Dutch words *tak* versus taak was 5 ms with a test segment duration of about 90 ms, i.e. 5.5%. The mean lengthening of our speaker's first syllables in two-word over one-word productions was 5.5 ms with a mean of 81.3 ms for the final-stress items, i.e. 6.76%, but 5 ms with a mean of 247.5 ms for the initial-stress items, i.e. 2.02%. (Two-word second syllables were lengthened by 2.24% in the final-stress case, but not at all in the initial-stress case, giving for the bisyllables as a whole 2.9% lengthening for final-stress items and less than 1% lengthening for initial-stress items.) Thus although some minimal first syllable lengthening occurred in initial-stress items, it may just have been too little for subjects to perceive.

It may be the case that in most English speech situations (where deliberately clear speech is not involved) syllabic lengthening as a boundary cue hardly exists. But we conclude that when durational cues to word boundaries are available, and are sufficient to exceed durational discrimination thresholds, listeners are able to exploit them.

ACKNOWLEDGEMENTS

This research was supported by the Alvey Directorate, UK, and by IBM UK Scientific Centre. We thank Brian Pickering, Francis Nolan, Brit van Ooyen and Kim Silverman for assistance.

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, pp. 152-178 (Cambridge University Press: Cambridge).

Butterfield, S., Cutler, A. (1988) *Segmentation errors by human listeners: Evidence for a prosodic segmentation strategy,* Proc. SPEECH '88, Vol. 3, 827-833.

Butterfield, S., Cutler, A. (1990) *Intonational cues to word segmentation in clear speech?* Proc. Inst. Acoust., (in press).

Cutler, A., Butterfield, S. (1989) *Natural speech cues to word segmentation under difficult listening conditions,* Proc. EUROSPEECH 89, Paris; Vol. 2, pp. 372-375.

Cutler, A., Butterfield, S. (1990) *Durational cues to word boundaries in clear speech,* Speech Comm., 9, (in press).

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

Cutler, A. and Norris, D.G. (1988) *The role of strong syllables in segmentation for lexical access,* J. Exp. Psy.: Hum. Perc. & Pert., 14, 113-121.

Lehiste, I. (1972) *The timing of utterances and linguistic boundaries,* J. Acoust. Soc. Amer., 51, 2018-2024.

Nooteboom, S.G., Doodeman, G.J.N. (1980) *Production and perception of vowel length in spoken sentences,* J. Acoust. Soc. Amer., 67, 276-287.

Taft, L. (1984) *Prosodic Constraints and Lexical Parsing Strategies,* PhD Diss., Univ. Massachusetts.

Umeda, N. (1975) *Vowel duration in American English,* J. Acoust. Soc. Amer., 58, 434-445.