PERCEPTUAL RELEVANCE OF ACOUSTICAL WORD BOUNDARY MARKERS

Hugo Quené

Institute of Phonetics, Utrecht University
the Netherlands

ABSTRACT

Measurements of segment durations in contrastive word boundary positions support the claim that acoustical boundary marking is realized differently among speakers. A tentative explanation of subjects' high accuracies in boundary detection experiments, viz. their ability to rapidly evaluate speaker-dependent boundary markers ("tuning in") was investigated. Ambiguous word pairs realized by 2 males and 2 females were presented either in lists of items realized by one speaker, or in a list of items of all 4 speakers randomized. From the results, it is concluded that subjects' simultaneous attention for multiple cues, rather than their "tuning in" to single cues, is responsible for the high boundary detection accuracy.

1. INTRODUCTION

Speech segmentation, the division of the continuous fluent speech signal into discrete words, is one of the most outstanding characteristics of human speech perception. Despite the numerous lexical ambiguities in the acoustic signal, listeners usually perform this task (as a preliminary for, or in interaction with word recognition) without much difficulty, although occasional errors do occur (2;7). Apparently, listeners are helped effectively by such top-down information as syntactic, semantic and contextual constraints, phonotactic restrictions of word structure, and sandhi phenomena. Besides, there are also bottom-up or acoustical phenomena related to word boundaries. Thus acoustically marked word boundaries in the fluent speech signal may help listeners in their segmentation. Among these phenomena, the following have been identified for English or Swedish by various researchers over the past decades as being functional in this respect: lengthening of pre-junctural consonant, aspiration of word-initial voiceless plosives, glottal stop or laryngealization of post-junctural vowels, and allophonic differences for /l,r/ in pre- or post-junctural positions (13;8;14).

In previous word boundary detection experiments (15), subjects were able to reach an overall accuracy of about 80% under conditions where no top-down information could have played a role (listening to ambiguous two-word sequences, not providing contextual or phonotactic cues). Contrary to e.g. (12;4), these results, as well as those by e.g. (10) show that listeners are able to make effective use of these acoustical word boundary markers as cues for speech segmentation, even without additional constraints based on top-down cues. Results of these previous experiments also suggest, that the following boundary marking phenomena had played a major perceptual role: (1) variation of word-initial vs. -final consonant allophone, (2) duration of ambiguous boundary consonant, (3) rise time of post-boundary vowel. Besides, (4) VOT of ambiguous plosives was observed to differ as a function of the intended boundary position.

Before establishing the perceptual relevance of these boundary markers more thoroughly, however, a rather unexpected finding from these experiments had to be further investigated, viz. the significant differences between speakers with regard to the produced acoustical (durational) boundary markers. Since such speaker effects (as well as language-specificity of these cues, as observed by (1)) have strong implications for the perceptual validity of the boundary markers mentioned above, we decided to investigate this matter first; the experiment reported here investigates subjects' ability to perceive speaker-dependent word boundary markers.

2. PRODUCTION

2.1. Material

Twenty-two word sequences were selected (2 word sequences with each of the 10 consonants /p,t,k,d,f,s,x,m,n,l,r/ which may occur word-initially as well as word-finally in Dutch, with this intervocalic ambiguous 'boundary consonant' in both word-final and word-initial position. (From (15) it was observed that word-final devoicing of /d/ did not affect boundary detection accuracy). The resulting 44 word sequences (11(consonants)x2 (sequences)x2(versions)) were embedded in sentences which disambiguated the word sequence.

2.2 Procedure

The 44 sentences were read aloud by 2 males and 2 females at a subjectively fast speech rate (to avoid pausing within sentences). Subsequently, the 4(speakers)x(44 word sequences)=176 ambiguous word sequences were spliced out of the original sentences by means of a computer programme (with visual and auditory feedback; sampling frequency 10 kHz; 12 bits resolution) and stored digitally.

Durations of the relevant speech portions (boundary consonant, VOT, and rise time of post-boundary V_2) were measured and analyzed.

2.3 Results

The data obtained show, that the four speakers produce the same durational difference between /C#/ and /#C/ boundary positions, for each of the three acoustical parameters under observation. However, the significance level of these differences is clearly speaker-dependent, as can be seen from Table I below.

Only one speaker, M2, produces significant differences for all three parameters in contrastive word boundary positions; the others produce some highly significant and some insignificant differences between /VC#V/ and /V#CV/ boundary positions.

Table I:
Resulting t-values (matched observations, pairwise deletion) of the durational differences between /C#/ and /#C/ boundary positions, for (1) duration of the ambiguous consonant, (2) VOT of ambiguous plosives, and (3) rise time of post-boundary vowel, for 4 speakers separately.

variable	M1	M2	F1	F2
(1)	- .132	-2.103^{*}_{*}	-2.694^{*}	-5.616^{***}
(2)	2.436_{***}	2.865_{*}	$.108_{***}$	1.899_{*}
(3)	4.806	2.229	5.483	2.103

*=p<.05; **=p<.01; ***=p<.001

In short, the durational observations suggest that acoustic marking of word boundaries is speaker-dependent. However, in (15) using a subset of these word sequences as stimuli, subjects obtained detection accuracies of over 75% with all four speakers. That is, although the acoustic marking of word boundaries is different among speakers, subjects were able to use these speaker-dependent markers to a considerable extent (as they were the only systematic cue). This high accuracy in boundary detection can therefore best be explained as a consequence of subjects' ability to evaluate these speaker-dependent acoustical cues rapidly, i.e. to "tune in" to them (analogously to feature adaptation in phoneme perception (5)).

In the following word boundary detection experiment, speaker-dependency in word boundary marking was further investigated. Presumably, it would be easier for listeners to "tune in" to speakers if they hear more stimuli realized by one speaker in a row, as compared to a listening situation in which they hear stimuli by several speakers in random order, and thus have to "tune in" to a different speaker for each new stimulus. This view constitutes the major hypothesis of the following detection experiment.

3. PERCEPTION EXPERIMENT

3.1. Design

In order to establish the effects of subjects' rapid evaluation of speaker-dependent word boundary marking, each subject had to perform boundary detection under two presentation conditions: (1) presentation of sub-lists of stimuli realized by the same speaker ('Sequences'), and (2) presentation of a randomized list of stimuli realized by all four speakers ('Randomized'). Thus, each subject had to perform boundary detection in all 176 word sequences (4(speakers)x22(word sequences)x2(boundary positions).

The relative order of these two presentation conditions was co-varied between subjects with the intended boundary position in an ambiguous word sequence (/VC#V/ vs. /V#CV/), yielding 2 different test tapes.

Four sub-tests were designed, with different internal ordering of the 4 same-speaker sequences, in order to neutralize interactions between the different same-speaker sequences. Thus, the experiment yielded 8 different test tapes, all containing the same 176 stimuli (word sequences) but different with respect to presentation condition and internal sub-ordering.

3.2 Material

The 176 digitized stimuli were DA-converted (10kHz; 12 bits) and re-recorded onto 8 separate audio tapes. Test items were preceded by 4 trial items and 10 filler items, and followed by another 10 filler items; trials and fillers were identical for all tapes.

The inter-stimulus interval was 3 sec; total time of each test tape was about 14 minutes.

3.3 Subjects and Procedure

Six (native Dutch speaking) subjects listened to each tape, yielding a total of (8x6=) 48 subjects. Most of them were undergraduate students in various Linguistics and Language studies. Their participation was voluntary, but they were paid a small amount (Hfl. 5,=) for their services. Subjects were assigned at random to one of the test tapes.

Subjects received written instructions, as well as a response booklet. For each stimulus item, this booklet gave two possible responses from which a forced binary choice had to be made. It was emphasized that they should not allow themselves to be influenced by the sometimes contrastive orthographies of the two possible responses (e.g. "zeis om" vs. "zij som").

Subjects listened to the test tape with closed headphones binaurally. Nine of them listened in a none-to-quiet room, the other 39 in a sound-treated booth. After the 4 trial items, the experimenter checked whether the instructions had been understood and the playback volume was comfortable, and gave additional oral instructions if necessary.

Responses agreeing with the boundary position as intended by the speaker were scored as 'correct', alternative responses as 'wrong'.

3.4 Results

Mean accuracy percentages for the various conditions are given in Table II below.

Table II:
Observed mean word boundary detection accuracies in percentages. Means for each bottom cell are calculated over 11(stimuli)x48(subjects)= 528 observations.

presentation	speaker	/VC#V/	/V#CV/	mean
	M1	91.8	66.4	
Sequences	M2	88.1	73.0	
	F1	87.0	73.0	
	F2	83.5	71.8	
	mean	87.6	71.0	79.3
	M1	93.8	71.0	
Randomized	M2	86.2	78.0	
	F1	90.5	74.5	
	F2	84.0	74.5	
	mean	88.6	74.5	81.6
mean		88.1	72.8	80.5

The dependent variable in the present experiment, viz. correct or wrong response, establishes a discrete (h.l. binary) random variable, following the binomial distribution with p=.5 and N=24 (subjects). However, since N.p>10, this distribution approximates the normal distribution so that the latter may be used as well (11).

Separate three-way analyses of variance were carried out with Speaker, Presentation and (intended) Boundary Position as main factors, integrating over subjects and words, respectively. From the resulting F-ratios, the minF' was calculated (3).

The SPEAKER variable yields an insignificant effect with minF'(3,82)<1. The same applies to the main variable which was of prime interest in this experiment, viz. PRESENTATION with minF'(1,3)=1.367 (insignificant). Thus, no significant difference in the proportion of correct responses (detection accuracy) could be observed between the two presentation conditions. Besides, the observed difference tends to be opposite to the prediction: subjects' word boundary detection is slightly more accurate in the Randomized condition as compared to the Sequences condition.

The only main factor yielding significance was BOUNDARY POSITION: minF'(1,13)=8.899; p<.025. As can be seen from Table II above, boundary detection accuracies were considerably higher in the /VC#V/ context as compared to those in the /V#CV/ context.

Significant interaction occurred between the factors Speaker and Boundary Position: minF'(3,90)=2.919; p<.05. Thus, detection accuracy between the two boundary positions (or contexts) was significantly different for the 4 speakers; the lowest difference was found for female speaker F2 (10.6%) and the highest difference for male M1 (24.1%). Other interactions did not reach significance.

4. DISCUSSION

Results of the present experiment show no significant effects of either Speaker nor Presentation. Although each of the four speakers under investigation employed to some extent different acoustical (durational) means to mark word boundaries in his (her) speech, these differences are not reflected in subjects' accuracy in word boundary detection. Listeners do not yield higher accuracy when listening to stimuli realized by one speaker to whom they could "tune in", as compared to the "Randomized" condition in which stimuli realized by four different speakers were presented.

These results allow for two possible explanations:

(a) Although word boundaries may be marked differently by different speakers, listeners pay simultaneous attention to all phenomena that may provide cues to word boundary location. That is, they do not focus on one acoustic cue which marks word boundaries for one speaker, switching attention to a different cue when stimuli realized by a different speaker are presented. Instead, listeners simultaneously focus on several phenomena which may or may not function to mark word boundaries, depending on who is speaking. Thus, they are "sensitive" to any of the cues the current speaker might possibly use. When switching to another speaker, they simply discard information provided by phenomena which do not help them, and rely more heavily on the phenomena which for this speaker assume the function of boundary markers.
Since all possibly relevant acoustical information for word boundary detection is monitored and evaluated continuously, the switching to different speakers has no effect on subjects' detection accuracy.

(b) The acoustical phenomena under investigation bear no perceptual relevance at all for word boundary detection. Although the four speakers realize significant differences for these acoustical markers (between /C#/ and /#C/ positions) to a different degree, these differences are perceptually irrelevant.
This interpretation of the results implies, that there are other acoustical cues, consistent between speakers, that systematically mark word boundary locations in fluent (Dutch) speech. These cues, yet unknown (but non-durational in nature), then have to be further investigated.

Since it is a quite common phenomenon that different acoustical cues simultaneously contribute to speech perception (as e.g. vowel length, VOT and F_0 all contribute to the voiced-voiceless distinction (9)), we feel that explanation (a) is the most likely. In a broader view, people generally use multiple cues to perceive significant aspects of their environment; our evaluation of other people, for example, is based on simultaneous impressions about their face, physical posture, what they say and how, and on their further behaviour. Probably, as in word boundary detection, numerous other (yet unknown) cues bear relevance as well. However, in order to accept explanation (a), we must disprove (b), i.e. it must be shown that the durational differences observed (viz. duration of the ambiguous intervocalic boundary consonant, and rise time of the post-boundary vowel) are perceptually relevant. If manipulation of these two parameters can be demonstrated to influence subjects' boundary detection, then explanation (b) must be discarded and (a) gains plausibility. Preliminary results suggest that this indeed seems to be the case; a more extensive study will be reported in the near future.

References

(1) BARRY, W.J. (1984) Perception of Juncture in English. In: M.P.R. van den Broecke and A. Cohen (eds.), Proceedings of the Tenth International Congress of Phonetic Sciences. Dordrecht/Cinnaminson NJ: Foris.

(2) BROWMAN, C.P. (1980) Perceptual, Processing: Evidence from slips of the ear. In: Fromkin (1980), pp. 213-30.

(3) CLARK, H.H. (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychological research, J. Verbal Learning and Verbal Behaviour, 12:335-59.

(4) COLE, R.A. and J. JAKIMIK (1980) Segmenting Speech into Words, J. Acoust. Soc. Am. 64(4):1323-32.

(5) COOPER, W.E. (1979) Speech Perception and Production: Studies in Selective Adaptation. Norwood NJ: Ablex.

(6) FROMKIN, V.A. (1980) Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand. London: Academic Press.

(7) GARNES, S. and Z.S. BOND (1980) A Slip of the Ear: A Snip of the Ear? A Slip of the Year?. In: Fromkin (1980), pp. 231-39.

(8) GÅRDING, E. (1967) Internal Juncture in Swedish. Traveaux de l'Institute de Phonetique de Lund, VI. Lund: Gleerup.

(9) HAGGARD, M., Q. SUMMERFIELD and M. ROBERTS (1981) Psychoacoustical and cultural determinants of phoneme boundaries: evidence from trading F_0-cues in the voiced-voiceless distinction, J.Phonetics 9:49-62.

(10) HARRIS, M.O., N. UMEDA and J. BOURNE (1981) Boundary perception in fluent speech, J. Phonetics 9:1-18.

(11) HAYS, W.L. (1973) Statistics for the social sciences. London: Holt, Rinehart and Winston. second edition.

(12) KLATT, D.H. (1980) Speech Perception: a model of acoustic-phonetic analysis and lexical access. In: R.A. Cole (ed.), Perception and Production of Fluent Speech. Hillsdale NJ: Lawrence Erlbaum Associates.

(13) LEHISTE, I. (1960) An acoustic-phonetic study of internal open juncture, Phonetica 5(suppl):5-54.

(14) NAKATANI, L.H. and K.D. DUKES (1977) Locus of segmental cues for word juncture, J. Acoust. Soc. Am. 62:714-19.

(15) QUENé, H. (1985) Word boundary perception in fluent speech: a listening experiment, Progress Report Inst. Phonetics Utrecht 10(2):69-85.