WORD SEGMENTATION IN MEANINGFUL AND NONSENSE SPEECH

Hugo Quené

Research Institute for Language and Speech, Rijksuniversiteit Utrecht
[quene@ruulif.let.ruu.nl]

ABSTRACT

This paper investigates the contribution of two phonetic word boundary markers on subjects' perceived word segmentation of ambiguous meaningful and nonsense word combinations. Both were realized in natural (both contrasting boundary positions intended) and synthetic speech (no boundary intended). Results show that markers are perceptually relevant, and that their contribution is the same for meaningful and nonsense stimuli. This suggests that phonetic markers are sufficient for word segmentation.

1. INTRODUCTION

In order to understand an utterance, its constituting words must be identified. To this end, a listener must (implicitly) locate the onset and offset points of words in an utterance. This *word segmentation* is argued to be a by-product of successful word recognition [1]. Listeners use stressed syllables as hypothetical word onsets. After recognition, a listener can anticipate on the subsequent word boundary and word onset (or attempt lexical access from the following stressed syllable, if the word is not yet recognized). However, this strategy only helps listeners in determining which syllables correspond to separate words. Word segmentation at the level of phonemes (or allophones), although necessary for word recognition, cannot be achieved through this strategy. Moreover, sensory word boundary information is indispensable in certain cases (e.g. words-within-words) [2].

In previous research [3,4,5], several acoustic-phonetic correlates of the (intended) word boundary have been identified. Roughly, two types of boundary phenomena can be discriminated: (1) 'explicit' boundary segments, e.g. laryngealisation or glottal stop (segmental, qualitative markers); (2) variations in segmental duration, e.g. word-initial consonant lengthening (durational, quantitative markers).

In this paper, it is examined whether these word boundary phenomena contribute to listeners' detection of word boundaries in connected speech, i.e. whether such phenomena are perceptually relevant. This can be investigated by means of manipulation of the boundary phenomena. If these markers are indeed perceptually relevant, then their manipulation should affect listeners' perceived word boundary position.

Given their (qualitative) nature, segmental boundary markers are less interesting for the present purpose. These "boundary segments" can only be perceived as the phonetic correlate of a (word or phrase) boundary. Hence, it is more interesting to assess the influence of *durational* cues on perceived word segmentation. This study concentrates on two such word boundary markers, viz. (a) the duration of the consonant adjacent to the word boundary, and (b) rise time of the vowel following the word boundary. In previous research [4,5], consonant duration was found to vary between 49 ms for intended /CVC#VC/, and 71 ms for intended /VC#CVC/; post-boundary vowel rise time varied between 19 ms and 13 ms, respectively [across 20 word combinations and 4 speakers].

In this study, the same type of stimulus material is used. Ambiguous two-word combinations may yield two distinctive sequences of two meaningful Dutch words (excluding segmentations involving geminates). For example, the combination /di(#)p(#)ɪn/ corresponds to the two Dutch two-word sequences **diep**

in "deep in" and **die pin** "that pin". In addition, ambiguous combinations yielding two *nonsense* words were included as stimuli. These enable a further test of the manipulated boundary markers: the perceptual relevance of the latter need not be limited to meaningful two-word combinations. The (unknown) intrinsic lexical effects on word segmentation are absent in nonsense word combinations.

In addition, even stronger evidence for the perceptual use of boundary markers can be obtained by using *synthetic* speech stimuli. Connected natural speech contains several acoustic-phonetic word boundary markers. The presence of all other (unmanipulated) boundary markers can be controlled in synthetic speech. If all other cues are absent, then any changes in the perceived boundary position between conditions can only be ascribed to manipulations of the phonetic word boundary markers.

In summary, the experiment reported here aims at providing evidence for the contribution of two durational boundary markers to listeners' perceived word segmentation, for combinations of either meaningful or nonsense words. These are realized (a) as /CVC#VC/ in natural speech [containing cues to the intended /C#/ boundary], (b) as /CV#CVC/ in natural speech [with cues to the intended /#C/ boundary], (c) in synthetic speech [containing no boundary cues]. In all realizations, boundary cues were manipulated by shortening and lengthening the durations of (a) the 'ambiguous' consonant adjacent to the word boundary, and (b) the rise time of the post-boundary vowel. Combining all conditions yields a 2x3x2x2 full factorial design.

2. EXPERIMENTAL METHOD

2.1. Stimulus material

Stimuli were constructed by combining a monosyllable (either a meaningful word or a non-word) having an ambiguous offset, with one having an ambiguous onset [e.g. /plat/ "plate", or /lat/ "late"]. Three types of boundary ambiguity were discriminated, depending on the number of intervocalic consonants and the possible positions of the word boundary:

Table I: Three types of word boundary ambiguity.

	nr.cons	ambiguity	
type 1	1	/V C#V/	/V #CV/
type 2	2	/VCC#V/	/VC#CV/
type 3	2	/VC#CV/	/V#CCV/

For all three types, the ambiguous boundary consonant could be either plosive, fricative or sonorant. However, type-3 word combinations with a sonorant boundary consonant are not allowed in Dutch. For each of the (3+3+2=) 8 remaining cells, 3 meaningful and 3 nonsense word combinations were constructed. For the meaningful (Dutch) word combinations, only monomorphematic words were used, and function words were avoided. Accent position was balanced between the two constituting words of a combination, and approximately balanced across boundary consonant categories and ambiguity types. For corresponding meaningful and nonsense combinations, the same member was accented.

In addition, 10 meaningful and 10 nonsense filler combinations were constructed, identical to the actual stimuli in all relevant aspects.

2.2. Natural speech material

Both contrasting versions of the 24 *meaningful* combinations were embedded in a meaningful sentence, which allowed only one segmentation. Corresponding sentences were as similar as possible with respect to number of syllables and words, stimulus position in sentence, etc. No important prosodic or syntactic break occurred within or immediately before or after the two-word sequence. In 9 out of 24 cases, it was necessary to extend the second of the two relevant words (in both versions) with a suffix, in order to fulfill these requirements. Both contrasting versions of the 24 *nonsense* combinations were embedded in a dummy carrier sentence, with a voiceless plosive immediately before and after the relevant sequence (for ease of excision).

All 2x2x24 stimulus sentences and 2x2x10 filler sentences were read twice by a professional speaker of Standard Dutch, and recorded on audio tape using high-quality equipment. Each two-word sequence was digitized (20 kHz sampling frequency, 9 kHz filtering, 12 bits) and excised from the carrier sentence [usually the second realization, unless affected by pausing, hesitation, mispronunciation, etc.]. Cuts were made at positive zero crossings (for nonsense sequences: after the preceding noise burst and within the following silent interval); no windows were applied. The resulting excerpts sounded natural, and did not suffer from

'clicks' at their onset or offset.

Natural sequences were processed identically to the (already processed) diphone source speech [6]. Digitized two-word sequences were fed into an LPC analysis (30 poles, window 25 ms, shift 10 ms). Subsequently, source type (voiced / unvoiced) and F_0 were established with a program using sub-harmonic summation and corrected if necessary. Filler combinations were digitized, excerpted and processed identically.

Subsequently, the analysis frames corresponding to (a) boundary consonant and (b) post-boundary vowel onset were established, by means of a segmentation program with auditory feedback and time-aligned displays of amplitude, voiced / unvoiced source, F_0, and original waveform. Vowel onset segments stretch from the first vowel frame to the frame with amplitude over 90% of the vowel peak amplitude (logarithmic).

2.3. Synthetic speech material

The 2x24 two-word sequences were generated by means of a diphone concatenation program [6]. In order to obtain the diphones used by this program, speech segments had been produced within (Dutch) nonsense words by the same speaker who realized the natural speech material in the present experiment (see above). From these utterances, the transition segments had been digitized (20 kHz, 12 bits), excerpted, and LPC-analysed.

The concatenation program was fed with phonetic transcriptions with accent symbols (no boundary symbols, "silence" phonemes or glottal stops). The output LPC analysis files (with marks for diphone and phoneme boundaries) were written to computer disk. The "accent" symbol yields a prominence-lending ('pointed hat') F_0 pattern on the appropriate vowel, superimposed on a declination line. After resynthesis, the diphone stimuli closely resemble natural stimuli. The crucial difference is that the synthetic speech does not contain any word boundary markers, since the diphones were originally realized word-internally. Again, filler combinations were input and concatenated identically.

Analysis frames corresponding to the two relevant intervals were established by the procedure described above, aided by the phoneme and diphone boundary marks in the LPC files. Vowel onset segments were not allowed to extend beyond the mid-vowel diphone boundary mark, nor beyond the F_0 turning point within the vowel (if accented).

2.4. Experimental conditions

The ambiguous boundary consonant and the post-boundary vowel onset were shortened (67%) or lengthened (134%) with regard to their original duration. Durations were manipulated by changing the number of samples for the appropriate frames [8]. Finally, the (2x2x (48+96)=) 576 manipulated two-word sequences, as well as the (2x2x10=) 40 unmanipulated fillers, were re-synthesised and stored on computer disk.

2.5. Stimulus tapes

The four stimulus conditions (meaningful--nonsense and natural--synthetic) were presented in separate blocks (pseudo-random order within blocks). Each block started and ended with 10 fillers. Four stimulus tapes were constructed, with counterbalancing between and within blocks. Tapes were recorded on DAT with 2.0 sec ISI (20 kHz, 9 kHz filter).

2.6. Subjects and procedure

Each tape was presented to 20 listeners (native Dutch, no reported hearing defects, language students) who received a small payment. They listened to the tapes over headphones (binaural) in a sound-treated booth. Their response sheet gave two possible responses (contrasting segmentations) for each stimulus; subjects were instructed to tick the appropriate one. Orthographic contrasts between responses were to be ignored. A short break was allowed between blocks on the stimulus tape.

Responses were fed (manually) into a computer, which calculated the rationalized arcsine [7] of the proportion of /#C/ responses (/V#CV/, /VC#CV/ or /V#CCV/, depending on combination type). The following section presents results of three tapes only, since remaining data are not yet available.

2.7. Results

The perceptual relevance of the manipulated boundary markers (a) consonant duration and (b) vowel rise time, should become apparent as a significant main effect of these factors. Influence of (c) the

speech source type and (d) the meaningful--nonsense difference on the perceptual relevance should become apparent as a significant interaction between these factors. In order to determine these effects, arcsine data were subject to an ANOVA with these four main factors. Two factors were added, viz. (e) the type of stimulus combination (8 types, fixed), and (f) the ambiguous combination (3 for each type, nested, random), see section 2.1. Main effects and relevant interactions are summarized in Table II; remaining interactions were all insignificant.

Table II: Summary of analysis of variance results.

	factor	F	df	p
(A)	cons.dur.	74.1	1,32	.001
(B)	vowel rise	20.5	1,32	.001
(C)	sp.source	149	2,64	.001
(D)	mean/nons	.4	1,32	n.s.
(E)	stim.type	.8	7,32	n.s.
(F)	stim.combi	28.3	32,1152	.001
AC		20.4	2,64	.001
AD		.1	1,32	n.s.
BC		6.1	2,64	.01
BD		1.4	1,32	n.s.
AE		3.0	7,32	.05
AF		2.0	32,1152	.01
CE		2.3	14,64	.05
CF		10.4	64,1152	.001
ACF		1.4	64,1152	.05

A Newman-Keuls post-hoc analysis on factor (C) showed a difference between natural stimuli, intended as /C#/, and both other speech source types ($p<.05$). Table III below illustrates the varying contribution of both boundary cues between the three source types (interactions AC and BC).

Table III: Mean percentage of /#C/ responses, for two manipulated boundary markers and three speech source types.

manipulation		NatC#	Nat#C	Synth
cons.dur.	Long	18	62	50
	Short	18	50	39
vowel ons.	Long	16	56	44
	Short	20	57	45

3. DISCUSSION

Both durational boundary markers under study are shown to contribute to word segmentation: manipulation affects subjects' perceived word boundary position. This perceptual relevance is identical in meaningful and nonsense conditions. Since the absence of anticipatory lexical information does not hamper word segmentation, phonetic cues seem to provide sufficient means to this end.

However, the natural speech conditions in Table III show that manipulations are only effective if they involve *post-boundary* markers (consonant in /#C/, vowel onset in /C#/). In addition, these data suggest that subjects pay primary attention to *un*manipulated markers in the natural stimuli, while the markers under study only play a secondary role. In general, subjects seem to perceive the intended boundary position on the basis of unmanipulated (probably segmental) cues. Durational cues contribute to this judgement, but only if the relevant speech segment follows the intended word boundary. The clustering of natural stimuli, intended as /#C/, with synthetic stimuli suggests that the latter also contain (uncontrolled) cues towards a /#C/ boundary position. Presumably, cues for *syllable*-initial position (in which the consonant diphones had been realized originally) were used for *word* segmentation in this experiment.

REFERENCES

[1] Cutler, A., & Norris, D. (1988) The role of strong syllables in segmentation for lexical access, *J. Experimental Psychology: Human Perception and Performance* 14 (1), 113-21.

[2] Frauenfelder, U.H. (1985) Cross-linguistic approaches to lexical segmentation, *Linguistics* 23, 669-87.

[3] Quené, H. (1987) Perceptual relevance of acoustical word boundary markers, *Proc. XIth Intl. Congress Phonetic Sc., Tallinn*, 6, 79-82.

[4] Quené, H. (1989) *The influence of acoustic-phonetic word boundary markers on perceived word segmentation in Dutch*. diss. Utrecht.

[5] Quené, H. (1991) Acoustic-phonetic cues for word segmentation. manuscript.

[6] Rijnsoever, P. van (1988) *From text to speech: User manual for Diphone Speech program DS*. IPO manual, 88.

[7] Studebaker, G.A. (1985) A "rationalized" arcsine transform, *J. Speech & Hearing Res.* 28, 455-62.

[8] Vogten, L.L.M. (1983) *Analyse, zuinige codering en resynthese van spraakgeluid*. diss. Eindhoven.