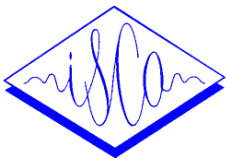

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

ON THE ABSENCE OF WORD SEGMENTATION AT "WEAK" SYLLABLES

Hugo Quené and Yvette Smits

Research Institute for Language and Speech, Rijksuniversiteit Utrecht,
Trans 10, 3512 JK Utrecht, the Netherlands
[quene@let.ruu.nl]

ABSTRACT

This research investigates the metrical segmentation strategy, which states that listeners attempt a lexical access at each metrically strong syllable, with this syllable as potential word onset. This paper reports on a "word spotting" experiment, where the target word corresponded with the second syllable of a two-syllable phrase. The onset of the target word was ambiguous. The metrical strength of the target word was manipulated by varying its phonological vowel length and accentuation independently. Neither of these two manipulations had a significant effect on subjects' hit rate or reaction time in spotting the target word. These results can be explained by assuming a different word segmentation strategy for Dutch, as compared to the metrical strategy reported for English. In addition, the results suggest that accentuation has an independent effect on word segmentation, most likely due to the enhanced perceptual salience of accented syllables.

10.21437/ICSLP.1992-47

1. INTRODUCTION

In order to understand a spoken utterance, a listener must identify the words which it contains. To this end, the speech signal should be segmented into discrete units, with which subsequent lexical access is attempted. Segmentation into word-size chunks seems to be the most efficient strategy, since this would result in units which can be used directly for subsequent lexical access. However, several authors [12, 7] have claimed a perceptual strategy, where listeners attempt a fresh lexical access at each strong syllable (i.e., a syllable with an unreduced vowel [5]); these strong syllables are used as potential word onsets. This would indeed be an efficient strategy, considering the fact that about 90% of word tokens begins with a stressed syllable [9, 3], which can never have a reduced vowel in English.

In addition, empirical evidence from "word spotting" experiments [6] supports this metrical segmentation strategy. In their crucial experiment, the target word (e.g. /mɪnt/) was followed by a VC syllable fragment with either an unreduced (e.g. /ɛɪv/) or schwa-like (e.g. /əʃ/) vowel. In the latter condition, the second syllable /təʃ/ is weak, hence the stimulus utterance is not segmented. The target word coincides with the onset of the (single) segmentation unit; this results in relatively short reaction times for the target word. In the former condition, the stimulus utterance is first segmented as /mɪn#tɛɪv/ (with strong syllable /tɛɪv/ as potential word onset). The target word is divided across these two segmentation units; this incorrect division produces longer reaction times in spotting the target word.

However, this experiment [6] provides only indirect evidence for the metrical segmentation strategy. The metrical segmentation strategy explains how listeners can determine word onsets in connected speech: metrically strong syllables are used as potential word onsets. This would imply that word recognition in connected speech is faster, if the target word begins with a strong syllable, as compared to words beginning with a weak syllable. However, in their experiment, the target word was not varied, but only the following syllable fragment was.

In the present experiment, we will seek more direct evidence for the metrical segmentation strategy, by varying the metrical strength of the critical target word itself. This amounts to an adapted replication "in reverse" of the experiment by Cutler and Norris [6]. In their experiment, the target word was contained in a two-syllable stimulus, with ambiguous syllable boundary (/CVCC.VC/ or /CVC.CVC/). Stimuli could have a "strong-strong" or "strong-weak" metrical pattern, which was varied by manipulating the nuclear vowel of the second syllable (e.g. /mɪntɛɪv/ vs. /mintəʃ/). The target word corresponded with the initial CVCC fragment, including the ambisyllabic consonant. Thus, the target word was always to be found at the *onset* of the stimulus sequence.

In the present experiment, the (Dutch) target word is contained at the *end* of a similar two-syllable stimulus (e.g. /xyplof/, target word underlined), with ambiguous syllabification as /CVC.CVC/ or /CV.CCVC/. The target word always corresponded with the final CVC fragment. The metrical pattern was varied in two ways, as will be explained below.

According to the metrical segmentation strategy, the following results are expected. If the stress pattern is "strong-weak", then there is no segmentation of the second syllable. All acoustic material is assumed to belong to the first (unrecognized) word, and the second syllable is not used as a starting point for lexical access (potential word onset). Hence, long reaction times (RT) are to be expected in spotting the word contained in the second syllable. If the stress pattern is "strong-strong", however, then the second syllable is indeed segmented. It is probably segmented as CVC: this follows from the delay in RT observed by [6] in their "strong-strong" condition, due to (incorrect) segmentation of the stimulus as /mɪn#tɛɪv/. Lexical access is attempted with the second CVC syllable, which in the present experiment does indeed correspond with the target word. Hence, short RTs are to be expected in this condition.

In English, vowel reduction is obligatory in unstressed syllables [8]. Hence, the occurrence of schwa-like vowels is a reliable acoustic correlate of the metrical structure of an utterance: weak syllables must be reduced. In Dutch, by contrast, this phenomenon is not obligatory: weak syllables need not be reduced, as exemplified by the first vowel in words like **tomaat** /to'mat/ "tomato", **metaal** /me'tal/ "metal" [8]. Consequently, vowel reduction (or the occurrence of schwa-like vowels) cannot be used experimentally to vary the metrical structure of a stimulus utterance in Dutch.

In the present experiment, we will therefore vary two other acoustic correlates of metrical structure. First, metrical structure depends partly on the respective weight of the syllables; in turn, this weight depends (partly) on the phonological length of the nuclear vowel [8]. Syllables ending in a long vowel and a one-consonant coda are "superheavy", while those ending in a short vowel and a one-consonant coda are "heavy". From a phonological viewpoint, the former are metrically stronger than the latter type of syllables [8]. Hence, the metrical structure of a stimulus utterance can be manipulated by varying the phonological length of its nuclear vowels.

Secondly, metrical structure can be affected by prosodic sentence structure. If a weak syllable is emphasized, it becomes metrically strong. In the present experiment, both methods of varying the metrical structure of the stimulus utterance will be applied orthogonally: the second syllable of a stimulus contains either a phonologically long or short vowel (e.g. /xyplof/ vs. /xyplɑ̄m/, target word underlined), and this syllable is either accented or unaccented.

2. METHOD

2.1. Design

As explained in the previous section, the two main factors in this experiment are the phonological vowel length (of the nuclear vowel of the second syllable), and the (presence or absence) of accent on the second syllable. Subjects' task is to spot the target word (similar to [6], which always corresponds with the final CVC fragment of a stimulus utterance. In addition, the test contains nonsense items (no target word contained in stimulus) and filler items (target word contained in other fragments of stimulus, e.g. first syllable). If the subject hears an existing word, he has to press a button first, and then say the recognized word. Responses based on recognized words other than the intended targets can thus be discarded. Dependent variables in this experiment are (1) the hit rate [percentage of detected target words], and (2) the reaction time in recognizing the target word.

2.2. Material

In order to construct the stimulus material, 33 pairs of CVC words were selected. The crucial contrast between the members of each pair was in the phonological length of the nuclear vowel (e.g. **loof** /lof/ "foliage" and **lam** /lɑm/ "lame"). Each member of a pair was prefixed with 2 different CVC non-word syllables, yielding 2x2x33= 132 two-syllable phrases. The combination of prefix and target syllables had to meet the following restrictions:

(a) The intervocalic consonant cluster must be a phonotactically legitimate Dutch word onset. Consequently, the second syllable has a phonotactically ambiguous onset (like many strong syllables in Dutch [9]), which prevents subjects from segmenting the stimulus phrase by means of phonotactical restrictions.
(b) The prefix CVC syllable may not be an existing Dutch word. Except for the target word, no other word may be contained in other parts of the resulting stimulus phrase (as would be the case in e.g. /niplef/ which also contains /ip/ "elm").

These 132 two-syllable phrases were entered (in phonetic transcription) into a Dutch speech synthesis system using concatenation of LP-coded diphones [11] (10 filter coefficients). All diphones had been excised from accented and unreduced syllables. Both segment parts constituting a diphone had been realised as part of the same syllable; hence the resulting speech does not contain syllable boundary cues.

For each phrase, two versions were synthesized, differing only in the presence or absence of an accent-lending F_0 movement during the second vowel ("pointed hat" pattern [13]). The prefix syllable was identical (in LPC parameter values) between phrases which differed only in second syllable, and/or in accentuation of that second syllable.

The above procedure yields 8 stimulus phrases for each pair of target words, as illustrated in Table I below. In total, 2x2x2x33=264 stimulus phrases were synthesized.

Table I: Overview of conditions, with example stimulus phrases. Bracketed numbers are for reference purposes only. In total, 33 of these octuples were synthesized as stimulus phrases. Quotes indicate a pitch accent on the following vowel.

		vowel in target word	
		long	short
prefix 1			
	+acc	x'ypl'of [1]	x'ypl'ɑm [5]
	-acc	x'yplof [2]	x'yplɑm [6]
prefix 2			
	+acc	l'ypl'of [3]	l'ypl'ɑm [7]
	-acc	l'yplof [4]	l'yplɑm [8]

In addition, 66 filler items were constructed, which contain a target word *not* corresponding with the final CVC fragment (e.g. /ternos/). Target words in filler items differed from all targets word contained in stimulus items. Likewise, 132 nonsense items were constructed, which did not contain an existing Dutch word.

Four stimulus tapes were constructed. In each tape, only two variants of each octuple were included, viz. those contrasting in Prefix, Vowel Length and Accent. Hence, for a particular octuple, the four tapes contained variants [1 and 8], [2 and 7], [3 and 6] or [4 and 5] respectively. All three factors were counter-balanced across tapes and across octuples. This yields 2x33 stimulus items on each tape. The 66 filler items and 132 nonsense items were identical between tapes. Hence, half of the items on each tape contained a Dutch word, while the other half did not. For each tape, all items were randomized.

All stimulus items, nonsense items and filler items were then resynthesized. For each stimulus utterance, the reference time from which reaction time was measured was established (i.e. the exact onset point of the target word). This was done by determining the onset of the second intervocalic consonant in the resulting sampled data speech file, by means of a speech editing program with visual (waveform) and auditory feedback.

Finally, stimulus tapes were constructed by DA-conversion of all utterances (at 10 kHz, 4.5 kHz low-pass filter, 12 bit) and subsequent recording on one channel of a stereo DAT. Reference signals were recorded on the other channel, synchronized with the predetermined onset points of the target word. Test items were separated by ISI's of 3 s.

2.3. Subjects and procedure

Five subjects listened to each tape. About half of the subjects had heard synthetic diphone speech previously; none of them reported hearing defects. Subjects listened individually to the speech channel of the DAT over binaural headphones in a sound-treated booth. They were instructed to press a button as soon as possible after they heard a Dutch word in the speech utterances, and to repeat the recognized word after this initial response.

The reference signal on the other channel of the DAT started a clock on a PC; this clock was stopped by subjects' response, and the reaction time was logged. Reaction times exceeding 2.5 s were ignored. Subjects' spoken responses were recorded on audio tape. If the responsed word differed from the intended target word (as judged by the second author), then the whole response was treated as a miss, and the corresponding RT was discarded. Responses with a correct initial consonant and nuclear vowel were treated as valid. In other words, for a stimulus item /xyplof/, responses like /los/ (with correct initial consonant and nuclear vowel) were treated as valid. On average, 18% of the responses on stimulus items were thus discarded.

2.4. Results

The results for both dependent variables, viz. (1) hit rate and (2) reaction time, are summarised in Table II below. It is clear that the hit rate, the percentage of correctly detected target words, does not differ between conditions of phonological vowel length and of accent. In other words, the percentage of spotted words is *not* higher for stimulus items in which the second (target) word contains a phonologically long vowel, or in which the target word is accented.

Table II: Absolute number of correctly detected target words (percentage in parentheses), with corresponding average reaction times in ms (standard deviation in parentheses). Data for each cell are based on 330 stimulus presentations [5 subjects x 66 stimuli].

	vowel in target word	
	long	short
+acc	n= 139 (42%) RT=1179 (396)	n= 149 (45%) RT=1108 (401)
-acc	n= 142 (43%) RT=1235 (443)	n= 146 (44%) RT=1256 (426)

Reaction time data were subjected to two analyses of variance. In the first analysis, results were collapsed across subjects. The main effect for factor Vowel Length was insignificant [$F(1,33)<1$], but factor Accent had a significant main effect [$F(1,33)=4.64$, $p<.05$]. In addition, considerable differences between stimulus items were observed, as shown by the marginally significant effect of factor Prefix (random, nested within Stimulus Phrase): $F(33,534)=1.41$, $.05<p<.10$. All other main and interaction effects were insignificant.

In the second analysis, collapsing results across stimulus items, a significant Subjects effect was observed (random, nested within Tape): $F(16,497)=4.17$, $p<.001$. The main effects for factor Vowel Length [$F(1,33)<1$] and Accent [$F(1,33)<1$] were both insignificant. The only interaction effect was between factors Tape and Accent [$F(3,16)=6.71$, $p<.01$]. All other main and interaction effects were insignificant.

If results from both analyses are combined into minF' ratios, then neither Accent nor Vowel Length has a significant effect. In summary, these results imply that phonological *vowel length* affects neither the frequency of occurrence nor the speed of recognition of the target word. The same applies to intonational *accent* of the target word. Nevertheless, accented words are recognized (and segmented) somewhat faster than unaccented words, although this effect was only found in one of the two analyses of variance.

3. DISCUSSION

The results presented above are inconsistent with the metrical segmentation strategy. As explained in the Introduction, this strategy would yield a higher hit rate and shorter reaction times in spotting the target word, if this word is metrically strong. However, the variation of two phonological correlates of metrical strength (of the target word) did not yield these predicted effects. This general outcome of the present experiment may be explained in two ways.

First, it could be possible that neither phonological vowel length nor accentuation is in fact related to the metrical structure of an utterance. Phonological theory claims than both factors affect metrical structure [8, 14, 1], but this relation may be absent in the perceptual domain. Hence, the two phonological factors manipulated in the present experiment, may be perceptually irrelevant for listeners retrieving the metrical structure of an utterance. However, this explanation seems to be implausible. Even if phonological evidence is ignored, the perceptual effects of varying metrical structure [6] and accent [10, 4] have been mutually consistent in previous research. This consistency illustrates the perceptual relation between accentuation and metrical structure.

The second explanation of the discrepancy in results between [6] and the present experiment, may be given by the differences between the two languages involved. In the former experiment, in English, both nuclear vowel reduction (full vs. reduced vowel) and accentuation were varied simultaneously between metrically strong and weak syllables. In the present experiment, in Dutch, phonological vowel length and accentuation were varied independently. Because vowel reduction and metrical structure are related in English (as explained in the Introduction), listeners may have used the former to retrieve the latter. No such strategy is possible in Dutch. The absence of any vowel length effect in our results indicates, however, that metrical structure by itself is not the decisive factor in word segmentation. Instead, the available evidence suggests that English listeners may rely on vowel quality in determining word onsets, while Dutch listeners may rely on other (yet unknown) phonetic cues to the same end.

In addition, varying accentuation of syllables within the stimulus phrases may have confounded the perceptual effects of metrical structure. Accented syllables are perceptually more salient than unaccented ones [10]; this increased salience may have triggered lexical access with the accented syllable in Cutler and Norris's experiment (where strong syllables were accented) as well as in the present experiment (where an insignificant difference between accent conditions was observed).

Finally, it should be noted that reaction times are conserably longer in the present experiment than those reported by [6]. In our opinion, this may be due to the synthetic nature of the stimulus speech in our experiment, which is considerably less intelligible than natural speech [2]. In addition, target words were always contained as the second syllable, which had an ambiguous onset. Hence, finding the onset of the target word was more difficult than in previous experiments, where they were contained at the onset of the stimulus items.

In conclusion, the results indicate that metrical structure (when defined in terms of phonological vowel length and accent) does not affect listeners' lexical access and word segmentation in Dutch. This suggests that in Dutch, metrical structure does *not* contribute to word segmentation and lexical access. However, the occurrence of an accented syllable (in Dutch and English) and of a syllable with unreduced vowel (in English) may trigger an attempted lexical access with this syllable as word onset. Using Occam's razor, these latter effects can be explained without reference to the metrical structure of the stimulus speech, viz. as a consequence of the enhanced perceptual salience of these syllables.

REFERENCES

[1] Baart, J.L.G. (1987) Focus, Syntax, and Accent Placement: towards a rule system for the derivation of pitch accent patterns in Dutch as spoken by humans and machines. dissertation Rijksuniversiteit Leiden.

[2] Bezooijen, R. van (1990) *Evaluation of speech synthesis for Dutch: Comparison of synthesis systems, intelligibility tests, and scaling methods.* Stichting Spraaktechnologie, Utrecht. SPIN-ASSP Report; 22.

[3] Cutler, A. & Carter, D.M. (1987) "The predominance of strong initial syllables in the English vocabulary", *Computer Speech and Language* **2**, 133-142.

[4] Cutler, A. & Clifton, C. (1984) "The use of prosodic information in word recognition", in *Attention and Performance. volume X: Control of Language Processes* (H. Bouma & D.G. Bouwhuis, eds.). Lawrence Erlbaum Ass., London. pp. 183-96.

[5] Cutler, A. & Fear, B. (1991) "Categoricality in acceptability judgements for strong versus weak vowels", in *Proceedings of the ESCA Workshop on Phonetics and Phonology of Speaking Styles*, Barcelona 1991, p. 18/1-5.

[6] Cutler, A. & Norris, D. (1988) "The role of strong syllables in segmentation for lexical access", *J. Experimental Psychology: Human Perception and Performance* **14**, 113-121.

[7] Grosjean, F. & Gee, J.P. (1987) "Prosodic structure and spoken word recognition", *Cognition* **25**, 157-188.

[8] Kager, R.W.J. (1989) *A Metrical Theory of Stress and Destressing in English and Dutch* Foris, Dordrecht. Linguistic Models; 13.

[9] H. Quené (1992a) "Integration of acoustic-phonetic cues in word segmentation", to appear in *The Auditory Processing of Speech: from sounds to words* (M.E.H. Schouten, ed.). Mouton De Gruyter, Berlin.

[10] Quené, H. (1992b) "Segment durations and accent as cues to word segmentation in Dutch". unpublished manuscript.

[11] Rijnsoever, P.A. van (1988) From text to speech: user manual for Diphone Speech program DS. IPO handleiding; 88. Internal publication, Inst. Perception Research, Eindhoven.

[12] Taft, L. (1984) Prosodic constraints and lexical parsing strategies. PhD thesis University of Massachusetts.

[13] 't Hart, J., Collier, R. & Cohen, A. (1990) *A perceptual study of intonation: an experimental-phonetic approach to speech melody.* Cambridge University Press, Cambridge.

[14] Visch, E.A.M. (1989) A metrical theory of rhythmic stress phenomena. dissertation Rijksuniversiteit Utrecht.