PHONOTACTIC AND ACOUSTIC CUES FOR WORD SEGMENTATION IN ENGLISH

Andrea Weber

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT

This study investigates the influence of both phonotactic and acoustic cues on the segmentation of spoken English. Listeners detected embedded English words in nonsense sequences (word spotting). Words aligned with phonotactic boundaries were easier to detect than words without such alignment. Acoustic cues to boundaries could also have signaled word boundaries, especially when word onsets lacked phonotactic alignment. However, only one of several durational boundary cues showed a marginally significant correlation with response times (RTs). The results suggest that word segmentation in English is influenced primarily by phonotactic constraints and only secondarily by acoustic aspects of the speech signal.

1. INTRODUCTION

Understanding spoken language requires the segmentation of a continuous speech signal into discrete words. This lexical segmentation problem can be solved by competition between a set of candidate words. Only those candidates which provide an optimal parse of the input win the competition [1]. However, previous research has shown that listeners can also use both acoustic and phonological cues, when available, to help solve the segmentation problem.

For example, the duration of speech segments varies at different positions in a word. Segments of word-initial syllables can be lengthened [see e.g. 2]; lengthening may thus provide a cue for a word boundary. Quené found that Dutch listeners can use durational cues to word boundaries when asked to choose between two alternative readings of an ambiguous two-word utterance [3]. However, acoustic cues to word boundaries are often small and variable [see e.g. 4]. In consequence acoustic cues (other than silence) may provide relatively weak assistance in segmentation.

Phonotactic constraints (restrictions on permissible sequences within syllables), on the other hand, can provide reliable boundary cues. For example, the sequence /ml/ is not legal within syllables in English; there must be a syllable boundary between /m/ and /l/. Because syllable boundaries are highly correlated with word boundaries, listeners could use their phonotactic knowledge for lexical segmentation by inserting a potential word boundary between such consonant sequences. Dutch listeners indeed find it easier to detect words in nonsense sequences when the word onsets are aligned with a phonotactic boundary (e.g., *rok*, 'skirt', in /fim.rɒk/) than when they are misaligned (e.g., *rok* in /fi.drɒk/; [5]). In /fim.rɒk/ a syllable boundary is required between /m/ and /r/, since /mr/ is not permissible within a syllable in Dutch. This leaves the onset of *rok* aligned with a syllable boundary. In /fi.drɒk/, however, a syllable boundary is required between /i/ and /d/, since /fid/ is not a possible syllable in Dutch due to final devoicing. This creates /drɒk/ in which the onset of *rok* is misaligned, and hence harder to detect.

Phonotactic constraints are not powerful enough to mark all word boundaries. In a corpus of continuous English speech only 37% of the word boundaries could be detected on the basis of phonotactic constraints [6]. However, when present, phonotactic cues, in contrast to gradient cues like segment duration, can reliably mark the onset of a word. Note that cues such as phonotactic constraints and acoustic differences do not replace the competition process but rather supplement and modify it.

The present study investigates the influence of both phonotactic and acoustic cues on the segmentation of spoken English. There are three parts to this study: (1) a word spotting experiment, investigating the influence of phonotactic constraints on lexical segmentation, (2) acoustic measurements investigating how speakers realize syllabification differences, and (3) correlation analyses of the established acoustic cues with the results of the word spotting experiment. In the word spotting experiment English listeners were presented with English speech stimuli. Their task was to detect any English word embedded in a list of nonsense sequences. The onset of the embedded word was either aligned with a clear syllable boundary (e.g., *luck* in /pun.lʌk/) or not (e.g., *luck* in /marflʌk/). In English /nl/ is not a legal consonant cluster within syllables and therefore the sequence /pun.lʌk/ requires a syllable boundary at the onset of the embedded word *luck*. On the other hand /fl/ is a possible syllable onset in English and the sequence /marflʌk/ does not require a boundary at the onset of *luck*, so that both /mar.flʌk/ and /marf.lʌk/ are possible syllabifications.

Previous research on phonotactic cues used a different manipulation [5]. McQueen contrasted the detection of embedded words when word onsets were aligned with a boundary (e.g., *rok*, 'skirt', in /fim.rɒk/) versus when they were misaligned (e.g., *rok* in /fi.drɒk/). If a speaker intended the word *rok,* it would never be misaligned in this way in Dutch. The manipulation of the present study ('alignment' versus 'no alignment') on the other hand is much more common. The predictions for the present study were however very similar to McQueen's study: embedded words were predicted to be detected faster in a word spotting task when the word onset was aligned with a clear syllable boundary according to English phonotactics (e.g., *luck* in /pun.lʌk/) than when it was not aligned (e.g., *luck* in /marflʌk/).

Since a sequence like /marflʌk/ allows two syllabifications, another point of interest was whether a speaker might use

acoustic cues to signal a word boundary in the absence of clear phonotactic alignment. Recordings were made with the speaker intending each syllabification of phonotactically ambiguous syllable boundaries (e.g., /mar.flʌk/ and /marf.lʌk/). Durational parameters of both intended syllabifications were measured. The durations of the segments /f/ and /l/ for example could be longer either when the segments were intended with a syllable boundary between them (/f.l/) or when they were intended as a cluster (/fl/). Once it had been established which acoustic phenomena vary systematically with the intended word segmentation in English, the question was whether listeners used these acoustic cues to locate word boundaries in the word spotting experiment. This was detemined by examining the correlation of acoustic measurements of the stimuli from the word spotting experiment and response times. If reliable acoustic cues to word boundaries are found in the stimuli and listeners use these cues for segmentation, then RTs from the word spotting experiment should correlate with the acoustic measurements.

2. PHONOTACTIC CUES

2.1 Methods

Subjects. Forty-eight native speakers of American English, students at the University of South Florida, were tested.

Materials and Procedure. 68 mono- and bisyllabic English nouns with initial /l/ (e.g. *luck*) or /w/ (e.g. *weapon*) were selected as target words. Each target word was appended to four different English nonsense syllables. The final consonants of two of the nonsense syllables aligned the onset of the following target word with a phonotactic syllable boundary (e.g., *luck* in /pun.lʌk/ and /fuʃ.lʌk/); the final consonants of the other two nonsense syllables did not align the onset of the target word (e.g., *luck* in /mɑɹflʌk/ and /dɔɪslʌk/). Different final consonants were used for the nonsense syllables within an alignment condition ('aligned' or 'not aligned'). Each target-bearing nonsense sequence contained its target word in final position but no other embedded word. In addition there were 55 filler nonsense sequences which contained embedded English words in final position with an initial consonant other than /l/ or /w/. A further 251 bi- and trisyllabic nonsense sequences contained no embedded words. Four lists were constructed. Each list contained all 306 filler sequences and 68 target-bearing sequences in a pseudo-random order, such that before each target-bearing sequence there was at least one filler that contained no embeddings. The fillers appeared in the same sequential position in all the lists. Each target also appeared in the same sequential position, but in only one of its possible contexts. Each list contained both types of target bearing sequences (target onset aligned and not aligned). 14 more representative practice items were added to the lists.

All materials were recorded onto DAT tape in a sound-proof booth by a female native speaker of American English. The speaker was instructed to avoid any clear syllable boundaries in the items for which two syllabifications were possible. Items were presented in the list orders using a portable computer and the NESU experiment control software. Subjects were instructed to listen to the nonsense sequences and press the button in front of them as fast as possible if they detected an embedded English word at the end of one of the nonsense sequences. They then had to say the word aloud. The computer timed and stored manual responses, and oral responses were recorded on tape. Each subject heard the 14 practice stimuli first, followed by one of the experimental lists. Prior to statistical analyses, RTs were adjusted so as to measure from the offset of the target words.

2.2 Results

Missed manual responses and manual responses that were accompanied either by no oral response or by a word other than the intended target word, as well as RTs outside the range of -200 to 2000 ms, were treated as errors. Seven target words with particularly high error rates were excluded from the analysis, leaving 61 words for the analysis. Mean RTs and error rates are given below in Table 1. Analyses of Variance with both subjects (F_1) and items (F_2) as the repeated measure were performed.

Measure	Aligned	Not aligned
RT	499	568
Errors	15%	22%

Table 1. Mean RTs in ms, measured from target offset, and mean percentage errors.

Three factor mixed ANOVAs were used, with in the subjects analysis experimental list as a between subjects factor, and initial sound (/l/ or /w/) and phonotactic alignment (with the two levels 'aligned' and 'not aligned') as within subjects factors.[1] A significant interaction (by subjects only) between initial sound and phonotactic alignment was found for the RTs ($F_1(1, 44) = 13.38$, $p = .001$; $F_2(1, 53) = 2.55$, $p > .1$).

Measure	initial sound	Aligned /pun.lʌk/ /jɑɹl.wɛpən/	Not aligned /mɑɹflʌk/ /mɔɪtwɛpən/
RT	/l/	518	554
Errors	/l/	13%	20%
RT	/w/	471	590
Errors	/w/	16%	23%

Table 2. Mean RTs in ms, measured from target offset, and mean percentage errors, split up by initial sounds.

ANOVAs were then performed separately for target words with initial /l/ and /w/. Mean RTs and error rates appear in Table 2. Phonotactic alignment significantly influenced RTs to words with both initial /l/ and initial /w/, though the effect was somewhat weaker for words with initial /l/ (/l/: $F_1(1, 44) = 7.09$, $p < .02$; $F_2(1, 30) = 4.58$, $p < .05$; /w/: $F_1(1, 44) = 32.56$, $p < .001$, $F_2(1, 23) = 12.23$, $p = .002$). Analyses of errors revealed very similar results. To summarize, the word spotting experiment showed clear evidence that segmentation in English

[1] In the items analysis, initial sound was a between items factor.

is influenced by phonotactic cues. Listeners found it easier to spot words that were aligned with a phonotactic boundary than words that lacked such alignment.

3. ACOUSTIC CUES

3.1 Methods

The target-bearing nonsense sequences from the word spotting experiment that lacked phonotactic alignment were recorded again by the same female speaker. She produced all sequences twice, with different intended syllable boundaries (e.g., /mar.flʌk/, /marf.lʌk/ and /dɔɪ.slʌk/, /dɔɪs.lʌk/).[2] Silent intervals (i.e. pauses) as a boundary cue were avoided, although silence would be a very strong boundary cue.[3] (Natural spoken utterances do not usually contain silence, but are rather continuous. If every word boundary were marked by a silent pause there would be no segmentation problem.) Using the Xwaves speech analysis software, several potentially relevant durations were measured. Since different final consonants were used for the nonsense syllables within an alignment condition, and embedded words started with either /l/ or /w/, most measurements did not apply to all stimuli. The duration of the first syllable vowel was measured for all items (122 pairs). VOT was measured for the clusters /pl/, /kl/, /tw/ and /kw/ (64 pairs). Fricative duration was measured for the clusters /fl/, /sl/ and /sw/ (58 pairs). Voiced duration of the /l/ was measured for the clusters /pl/, /kl/, /fl/ and /sl/ (68 pairs). Since /l/ is often partially devoiced after an aspirated stop, the duration of that part of the /l/ that was voiced was also measured in the clusters. These measures which differed significantly between the two intended syllable boundaries were then also measured in the stimuli from the word spotting experiment.

3.2 Results

Mean durations of the acoustical measurements for the productions with differing intended syllabifications are given below in Table 3. Two factor mixed ANOVAs were used for each of the four measurements, with consonant cluster as a between items factor, and intended syllabification (with the two levels 'single onset' and 'onset cluster', e.g. /mɑɹf.**l**ʌk/ and /mɑɹ.**fl**ʌk/) as within items factor.

Highly significant effects of intended syllabification and no interaction with consonant cluster were found in the vowel duration ANOVA and in the voiced duration of /l/ ANOVA. There was no significant effect of intended syllabification in the VOT ANOVA. Since in the ANOVA for fricative duration there was a significant interaction between intended syllabification and consonant cluster, separate ANOVAs were performed for the three consonant clusters containing fricatives. Significant effects of intended syllabification were found for the clusters /fl/ and /sl/ (both with fricatives longer in single onset condition), but not for /sw/. In summary, three durational measurements, first syllable vowel duration, voiced duration of /l/ and fricative duration were found to vary systematically with the intended syllabification.

Measure	single onset /mɑɹf.**l**ʌk/ /mɔɪt.**w**ɛpən/	onset cluster /mɑɹ.**fl**ʌk/ /mɔɪ.**tw**ɛpən/
First syllable vowel duration (all items)	150	177
VOT (stops)	85	79
fricative duration (fricatives)	153	136
voiced duration /l/	44	24

Table 3. Mean durations in ms for the two intended syllabifications.

When RTs from the word spotting experiment to sequences with no phonotactic alignment were split up by clusters, the effect size of phonotactic alignment differed across consonant clusters. RTs to words with initial /l/ were, for example, especially fast if they followed /s/ (e.g., *lift* in /mɔɪslɪft/) compared with /p/ (e.g., *lift* in /jiplɪft/, mean difference of 198 ms), even though /sl/ and /pl/ are both possible syllable onsets in English and therefore share the same phonotactic status. These differences in effect size might be explained by acoustic cues to syllabification in the speech signals. If the speaker (unintentionally) marked boundaries in the word spotting stimuli with durational cues, then listeners may have used those cues for segmentation.

Correlation analyses for duration measurements (of the speech signals used for the word spotting experiment) with RTs were performed for the three measures in Table 3 which showed a significant difference. The first measure, 'vowel duration of the first syllable', failed to show a significant correlation with RTs in the word spotting experiment, though it was established as an acoustic difference a speaker may use to signal a boundary. The reason why no correlation was found for vowel duration could lie in the parameters chosen for the measurements. When the vowel in the nonsense syllables was followed by the approximant [r] (51 times out of 122), vowel duration was measured including the approximant since the speech signal does not show a clear ending of the vowel. Consequently vowel durations differed noticeably depending on whether they included a following approximant or not. Similarly the second measure, 'fricative duration', failed to show a significant correlation with RTs, though it was also established as an acoustic difference a speaker may use to signal a boundary. The third measure 'voiced duration of /l/' showed a marginally significant negative correlation with RTs when all clusters containing /l/ were included in the analysis ($r(68) = -.23$, $p = .058$). When only the two clusters with the most extreme mean RTs and the biggest difference in the acoustic measurement (/sl/ and /pl/) were included in the analysis, the correlation was fully significant ($r(46) = -.31$, $p < .04$). Longer voiced duration of /l/ marked /l/ as syllable initial for the listeners, which in

[2] The intended boundary can only be varied where phonotactics do not force a boundary (e.g. /mɑɹflʌk/), otherwise the speaker would have to produce onset clusters that do not occur in his or her language (e.g., /nl/ in */pu.nlʌk/).

[3] The closure portion preceding the burst of a stop was not regarded as a pause.

consequence marked the onset of the embedded word as aligned and made it easier to spot. The results suggest that at least some acoustic information, when available, contributes to the process of segmentation.

4. DISCUSSION

Earlier studies have shown that the process of spoken word segmentation is influenced by phonological information. Dutch listeners find it easier to detect words in nonsense sequences when the words are aligned with a phonotactic boundary than when they are misaligned [5]. Language-specific metrical information also appears to provide listeners with important segmentation cues. In a stress-timed language like English, the majority of content words begin with strong syllables [7]. Cutler and Norris found that listeners rely on strong syllables in English to initiate the lexical search [8]. Word boundaries however, cannot always be derived unambiguously from phonotactic or metrical information. Acoustic boundary markers have also been found to influence lexical segmentation in previous research. Dutch listeners can use durational cues to word boundaries when asked to choose between two alternative readings of an ambiguous two-word utterance [3].

The results of the present study support the claim that the legality of phoneme sequences is used to help solve the segmentation problem. English subjects, when asked to spot embedded words in nonsense sequences, found it easier to spot words that were aligned with a English phonotactic boundary than words that lacked such clear alignment. This effect was found for both response times and error rates. In nonsense sequences that lacked phonotactic alignment, however, differences in effect size were found across different consonant sequences. Acoustic boundary cues that align or misalign word onsets might have been responsible for this variability in the RTs.

Acoustic analyses of nonsense sequences with two intended syllabifications were made, in order to establish which durational parameters a speaker may use to signal word boundaries in the absence of phonotactic alignment. Vowel duration of the first syllable, fricative duration and voiced duration of /l/ were found to vary systematically with the intended syllabification in recorded nonsense sequences. These three durational parameters were then measured in the speech signals of the word spotting experiment stimuli. However, only one durational measurement, namely voiced duration of /l/, correlated with the RTs from the word spotting experiment. Subjects' perceived word segmentation might still have been affected by other acoustic boundary markers not measured in this analysis, but the parameters measured here represent the most likely options. In general, acoustic cues may be too variable and too small to be used extensively by listeners for lexical segmentation. Thus different boundary cues could provide different degrees of assistance in segmentation. Phonotactic boundaries may be more powerful since they are reliable when present, whereas the observed durational cues may be less powerful segmentation cues since they are gradient and speaker and speech rate dependent.

In conclusion, the results of the present study show that listeners use phonotactic information to identify likely word boundaries in continuous English speech. Acoustic aspects of the speech signal played only a small role in the observed effect.

5. ACKNOWLEDGEMENTS

This research was supported by the Max Planck Institute for Psycholinguistics and by NIDCD 00323. I thank Anne Cutler, Winifred Strange, James McQueen and Natasha Warner for their comments on this work. Please address all correspondence to: Andrea Weber, Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands. Email: andrea.weber@mpi.nl.

6. REFERENCES

1. McQueen, J.M., Norris, D.G., and Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 621-638.

2. Gow, D.W., and Gordon, P.C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 344-359.

3. Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics*, **20**, 331-350.

4. Nakatani, L.H., and Dukes, K.D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, **62**, 714-719.

5. McQueen, J.M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, **39**, 21-46.

6. Harrington, J., Watson, G., and Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. *Computer Speech and Language*, **3**, 367-382.

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

8. Cutler, A., and Norris, D.G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113-121.