SYLLABIC CUES TO WORD SEGMENTATION

Cecilia Kirk
University of Massachusetts, Amherst, U.S.A.

ABSTRACT
This research supports the claim that listeners use their knowledge of what constitutes a well-formed syllable in their language to select the most likely locations for word boundaries. In a word-spotting experiment, English listeners found it much more difficult to spot words when the speaker’s intended syllabification, as determined by allophonic variation, was misaligned with the word boundary. When there were no obvious word boundary cues, English listeners adopted parsing preferences that maximized syllable onsets and attracted consonants to stressed syllables. Both types of segmentation strategies can be accounted for by the Possible Word Constraint (PWC) [1] which penalizes any parse of the speech signal which leaves an impossible word as a residue.

1. INTRODUCTION
The Shortlist model of spoken word recognition proposes that lexical competition is the core mechanism for segmenting continuous speech, but that this competition process can be modulated by word boundary information [1],[2],[3],[4]. When segmenting the speech stream listeners do not assume that words can begin at any point in the input, but instead they use a variety of cues such as the phonotactics [4] and metrical structure [5] of their language to identify likely word boundaries. A unified account of the different segmentation cues used by listeners has been proposed by the Possible Word Constraint (PWC) [1] which claims that listeners use their knowledge of what constitutes a possible word in their language to limit the number of competing candidate words. The present study uses the wordspotting task to investigate the role of the syllable in the segmentation of connected speech in English for those cases where syllabification is not dictated by phonotactics. I investigate two types of syllabification strategies that might be used when parsing the acoustic input: those used to parse the input when syllabification is unambiguously determined by allophonic cues, and those that might come in to effect when syllabification of the input is not determined by either phonotactic or allophonic cues.

In current theories of lexical access, words are represented as concatenations of segments without any structure internal to the word. This is clearly inadequate, if it can be shown that listeners use their knowledge of different allophonic variants of a phoneme to determine the likelihood of a word boundary. It is hypothesized that listeners will find it difficult to spot words when the speaker’s intended syllabification, as indicated by allophonic variation, does not coincide with the word boundary. Two different versions of this hypothesis were tested: one involving retroflexion of alveolar stops, and the other involving glide devoicing. In English, when the consonants /dr/ are surrounded by vowels, they can be parsed either together in the syllable onset, or with a syllable boundary between them. When an alveolar stop occurs together with /t/ in the syllable onset, it is very commonly pronounced with a strongly retroflexed articulation. Retroflexion of the alveolar stop can provide important information about segmentation possibilities. If the listener is able to extract and make use of allophonic information to parse the incoming signal, then the word “rock” should be more difficult to spot when the /d/ in /vudrak/ is retroflexed than when it is not. Likewise, devoicing of a glide following an aspirated stop gives an important cue to a speaker’s intended syllabification, since this only occurs in syllable onsets. If listeners are able to make use of syllabic information to reduce the number of lexical competitors, then the word “wine” embedded in /zikwayn/ should be more difficult to spot when the [k] is syllabified together with the [w] in the onset than when the [k] is syllabified as a coda and no aspiration or devoicing occurs.

Even in the absence of positive cues to syllabification, it is possible that listeners assign structure to the incoming acoustic signal. Current models of spoken word recognition predict that “this moon” and “the smooth” will be segmented equally easily. However, if listeners follow a strategy of maximizing the syllable onset when segmenting the input, ease of segmentation will not be equivalent in both instances. In this study, I test the prediction that the word “snake” will be spotted more quickly and accurately in [visneyjk] than spotting of the word “neat” in [visnit]. Wordspotting in the latter case will be more difficult since maximization of the syllable onset results in the parse [vi.snit] where the syllable boundary is misaligned with the word boundary, even though there is no phonotactic requirement that [s] and the following nasal be syllabified together. It is further hypothesized that English listeners use a segmentation strategy whereby intervocalic consonants which are otherwise syllabically ambiguous are parsed together with an adjacent stressed syllable. This predicts that listeners will find it easier to spot the word “east” when embedded in [slam.ist] than when embedded in [sla.mist]. The two parsing strategies of onset maximization and stress attraction will be referred to in the remainder of the paper as grammatical constraints on syllabification since they are determined by the grammar in the absence of positive cues to syllabification. However, by assuming this label, I do not
wish to imply that phonotactic knowledge and allophonic knowledge are not part of the speaker’s grammar.

2. METHOD

2.1. Participants

Fifteen students from the University of Massachusetts participated in the experiment in exchange for course credit. All were native speakers of American English and none reported any hearing or speech disorders.

2.2. Materials

All targets were English monosyllabic content words and appeared in the final position of bisyllabic nonsense strings. All vowels in the nonsense residue (i.e. the prefix) were tense vowels.

Retroflexion. The twenty-two targets consisted of the shape rVC(C) (e.g. run). Each of the targets was preceded by a prefix of the shape CVD, where D represents either a voiced or voiceless alveolar stop. In half of these items, the alveolar stop at the end of the prefix was pronounced strongly retroflexed. Each target from the retroflexed list was paired with a target from the unretroflexed list which was matched for frequency and shared the same CVD as the nonsense prefix.

Glide devoicing. All twenty-four targets were of the form wVC(C) (e.g. wine). The prefix was of the shape (C)CVk. In half of these items, the voiceless velar stop at the end of the prefix was pronounced strongly aspirated and the initial glide of the target word was devoiced. Each target from the aspirated/devoiced list was paired with a word from the unaspirated/voiced list which was matched for frequency and shared the same (C)CVk as the nonsense prefix.

Maximize the onset. Half the twenty-four targets consisted of sNVC (e.g. snake) and the other half consisted of NVC (e.g. neat) where N represents either a labial or alveolar nasal consonant. Each word in the sNVC list was paired with a word the NVC list. The paired items were matched for frequency and shared the same CV as the nonsense prefix. The prefixes for words from the NVC list had an additional “s” in final position.

Stress attracts consonants. All twenty targets began with a vowel. The prefix consisted of CVR, where R represents either a labial or alveolar nasal, or a lateral. Half the bisyllables were pronounced with primary stress on the initial syllable and the other half with primary stress on the second syllable. Each of the targets from the list with initial stress was matched with a target of similar frequency from the list with final stress. The paired items also shared the same CVR prefix.

The materials used to test the hypothesis that listeners use allophonic variation to select likely word boundaries as well as the hypothesis that listeners maximize onsets in the absence of positive boundary cues were tested together in a single block of trials. This was followed by a second block of trials which tested the hypothesis that stress attracts consonants. Only twelve of the fifteen listeners heard the second block of trials. Before the first block of trials, listeners were presented with a list of fourteen bisyllabic practice items. In six of these practice items, either the first or second syllable was a word. Since all the target words were in final position, there was concern that listeners would develop strategies causing them to disregard the prefix and focus only on the second syllable of each item. To discourage such strategies, the first block of trials included twenty bisyllabic distractor strings which contained a word embedded in the initial syllable. 174 bisyllabic nonsense strings which contained no embedded real words were used as fillers. Not every listener heard all fillers, since a different experimental file was constructed for each listener. In the second block of trials, there were thirteen bisyllabic practice items, five of which had words embedded in either initial or final position. The second block contained ten distractors and fifty fillers.

Table 1. Mean percentage missed targets (Errors) and mean reaction time for correct detection (RT, in ms), measured from target-word offset.

<table>
<thead>
<tr>
<th></th>
<th>Aligned</th>
<th>Misaligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retroflexion:</td>
<td>[vud.rak]</td>
<td>32%</td>
</tr>
<tr>
<td>glide devoicing:</td>
<td>[zik.wayn]</td>
<td>17%</td>
</tr>
<tr>
<td>maximize the onset:</td>
<td>[vi.sneyk]</td>
<td>33%</td>
</tr>
<tr>
<td>stress attracts:</td>
<td>[slám.ist]</td>
<td>6%</td>
</tr>
<tr>
<td>RT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retroflexion:</td>
<td>[vud.rak]</td>
<td>739</td>
</tr>
<tr>
<td>glide devoicing:</td>
<td>[zik.wayn]</td>
<td>719</td>
</tr>
<tr>
<td>maximize the onset:</td>
<td>[vi.sneyk]</td>
<td>695</td>
</tr>
<tr>
<td>stress attracts:</td>
<td>[slám.ist]</td>
<td>473</td>
</tr>
</tbody>
</table>
2.3. Procedure

The materials were recorded by a male speaker of American English. Every bisyllable was pronounced with compound stress, with the exception of the subset of items and fillers used to test the hypothesis that stress attracts consonants. Of the items used to test whether stress attracts consonants, half were pronounced with a stress pattern of strong-weak and the other half with the stress pattern of weak-strong. Each listener heard the items presented in a different pseudo-random order such that there was at least one filler between two target-bearing items. Items were presented over headphones with an inter-item interval of 5 seconds. Listeners were tested individually in a sound-protected room. They were told that they would hear a list of nonsense bisyllables and that their task was to try to spot real words embedded at either the beginning or end of some of the bisyllables. They were asked to press a button as fast as possible if they spotted a word and then to say aloud what that word was. Responses were measured from the onset of each target-bearing bisyllable.

3. RESULTS

All button-press responses associated with incorrect or missing oral responses were treated as errors. Reaction Times (RTs) were adjusted so as to measure from target word offset by subtracting the duration of the complete bisyllable from the raw RT for that item. Correct responses that were more than two standard deviations from the overall mean were also treated as errors. Mean error rates (percentage of targets missed) and mean RTs (measured from target offset) are shown in Table 1. Analyses of variance (ANOVAs) were performed on errors and RTs, with both subjects (F1) and items (F2) as the repeated factor. The mean RTs for correct responses and mean error rates are given in Table 1.

3.1. Allophonic constraints on syllabification

Listeners were much slower and less accurate at spotting words when the speaker's intended syllabification, as determined by allophonic variation, did not coincide with the word boundary.

Retroflexion. When the word boundary was marked by retroflexion (e.g., “rock” in [vud rak] versus “rock” in [vud rak]), listeners found it more difficult to spot the target word than when there was no retroflexion. There was a highly significant effect of errors (F(1, 14) = 24.87, p < 0.000, MSE = 2.71; F(2, 10) = 8.18, p < 0.05, MSE = 11.24) and RTs (F(1, 14) = 11.43, p < 0.005, MSE = 44466; F(2, 10) = 9.21, p < 0.05, MSE = 87030).

Glide Devoicing. One of the prefixes was perceived as a word by a number of listeners and one of the targets was not correctly identified by any listener. These items were removed from the analysis which reduced the number of targets per condition to ten. When the word boundary was masked by aspiration of the consonant preceding the target and devoicing of the glide in target-initial position (e.g., “wine” in [zi.k²wavn] versus [zik.wavn]), listeners found it more difficult to spot the target word than when these cues to syllabification were absent. There was a significant effect of errors (F(1, 14) = 91.69, p < 0.000, MSE = 1.44; F(2, 9) = 13.63 p < 0.005, MSE = 14.56). RTs were significant only by items (F(2, 9) = 5.33, p < 0.05, MSE = 31254). The lack of significance by participants is due to the fact that three of the fifteen subjects correctly identified only a single target in the misaligned condition and in all three cases their RT for this single item was well below the overall subject mean for this condition.

3.2. Grammatical constraints on syllabification

Listeners found it more difficult to spot words when the syllable boundary, as determined by the syllabic parsing preferences of onset maximization and stress attraction, did not coincide with the word boundary.

Maximize the onset. Three of the prefixes were perceived as words by a number of participants. Items with these prefixes were removed from the analysis which reduced the number of targets per condition to nine. When the syllable boundary, as determined by onset maximization, was misaligned with the word boundary, listeners were much slower and less accurate at spotting the target word than when the word and syllable boundaries were aligned (e.g., “neat” in [vi.snit] versus “snake” in [vi.snety]). There was a significant effect of errors (F(1, 14) = 73.50, p < 0.000, MSE = 0.80; F(2, 8) = 11.20, p < 0.01, MSE = 8.75) and RTs (F(1, 14) = 7.51, p < 0.05, MSE = 49845; F(2, 8) = 20.13, p < 0.005, MSE = 20879).

Stress attracts consonants. One of the prefixes was perceived as a word by a number of participants and so was removed from the analysis reducing the number of targets per condition to nine. Listeners found it difficult to spot words when the target item was stressed since this appeared to encourage the final consonant of the prefix to be syllabified as the onset of the target, (e.g., “east” in [sla.mist] versus “east” in [slam.ist]). There was a significant effect of errors (F(1, 11) = 51.28, p < 0.000, MSE = 1.95; F(2, 8) = 13.66, p < 0.05, MSE = 9.76) and RTs (F(1, 11) = 6.23, p < 0.05, MSE = 9276; F(2, 8) = 14.38, p < 0.005, MSE = 12688).

4. DISCUSSION

I have argued that words are perceived not just as linear sequences of phonemes, but that they have internal structure that is exploited in online processing. When interpreting sequences of phonemes, listeners appear to use grammatical constraints specific to their language to
structure the input, thus limiting the number of competitors under consideration at any one time. When the speaker’s intended syllabification is unambiguously marked by allophonic variation, this information is used in the online segmentation of speech. When syllabification is neither phonotactically or phonetically determined, listeners still assign structure to the input. Preferred syllables shapes are those where onsets are maximized and consonants syllabify with adjacent stressed syllables.

Although word boundaries and syllable boundaries are not perfectly correlated, using syllable onsets to posit the location of possible word onsets limits the number of hypotheses about competing candidate words. The Possible Word Constraint (PWC) crucially depends on the listener’s ability to detect syllable boundaries, penalizing candidates with mismatched word and syllable boundaries. The results presented here suggest that syllable boundaries are being assigned to the incoming speech signal, even in the absence of obvious cues to syllabification. Under the assumption that the input is being structured in this way, the PWC provides a unified account of phonotactic and allophonic cues to segmentation as well as segmentation preferences, such as onset maximization. Under this model, activation of candidate words is reduced when their edges are misaligned with phonotactically, allophonically, or grammatically determined syllable boundaries causing listeners to avoid parses of continuous speech that result in impossible words.

5. ACKNOWLEDGEMENTS
This research was supported by Grant DC01708 from NIH to John Kingston. Correspondence should be sent to Cecilia Kirk, Linguistics Department, South College, University of Massachusetts, MA 01003, U.S.A. Email: cjkirk@linguist.umass.edu

6. REFERENCES