Segmenting Ambiguous Phrases Using Phoneme Duration

Keren B. Shatzman

Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
Keren.Shatzman@mpi.nl

Abstract

The results of an eye-tracking experiment are presented in which Dutch listeners' eye movements were monitored as they heard sentences and saw four pictured objects. Participants were instructed to click on the object mentioned in the sentence. In the critical sentences, a stop-initial target (e.g., "pot") was preceded by an [s], thus causing ambiguity regarding whether the sentence refers to a stop-initial or a cluster-initial word (e.g., "spot"). Participants made fewer fixations to the target pictures when the stop and the preceding [s] were cross-spliced from the cluster-initial word than when they were spliced from a different token of the sentence containing the stop-initial word. Acoustic analyses showed that the two versions differed in various measures, but only one of these – the duration of the [s] – correlated with the perceptual effect. Thus, in this context, the [s] duration information is an important factor guiding word recognition.

1. Introduction

In order to understand a spoken utterance, the continuous speech signal must be segmented into lexical units. Current models of spoken-word recognition such as TRACE [1], Shortlist [2], and DCM [3] agree that, as speech unfolds over time, words that are fully or partially consistent with the input become activated and compete among one another. The outcome of the competition is that the input is parsed into a sequence of non-overlapping words.

The lexical competition process is determined, to a large extent, by the information in the acoustic signal. An important line of research has therefore focused on identifying acoustic cues that may mark word boundaries. One consistent finding in the acoustic-phonetic literature is that phoneme duration varies as a function of the position of the phoneme with respect to word boundaries. Thus, for example, phonemes in word-initial position tend to be longer than in word-medial or word-final position [4, 5, 6].

The influence of phoneme duration on segmentation has been demonstrated in a study with ambiguous two-word sequences, such as the Dutch phrases *diep in* ("deep in") and *die pin* ("that pin") [7]. Using a forced-choice task, it was shown that Dutch listeners made use of the duration of the intervocalic consonant in segmenting these word pairs. The study showed that manipulating the duration of this intervocalic consonant influenced listeners' explicit lexical segmentation judgements.

Other studies have shown the influence of phoneme duration on segmentation using on-line measures. A study in French [8] examined segmentation in a liaison environment, that is, in phrases where resyllabification occurs across word boundaries. In contexts like *dernier oignon* (last onion), the final [ʁ] of *dernier* is produced and resyllabified with the following syllable, making the phrase homophonous with *dernier rognon* (last kidney). The results of the study suggest that French listeners' segmentation of an ambiguous liaison phrase (e.g., *dernier oignon / rognon)* is influenced by the length of the liaison consonant: the consonants in the liaison environments were shorter than the word-initial consonants (e.g., [ʁ] in *dernier oignon* vs. *rognon*). Similarly, in an English study [9], evidence was found for priming of words in two-word sequences (e.g., of *lips* in *two lips*) but not when the words were pronounced as part of single-word sequences (e.g., *tulips*). Given the fact that the word-initial consonants (e.g., the [l] in *two lips*) were longer than the non-initial consonants (e.g., the [l] in *tulips*), the authors concluded that listeners were using this acoustic marker of word onset in lexical access and segmentation. However, in both [8] and [9], it was not demonstrated that the duration of the critical consonants is what affected listeners' segmentation. In other words, the link between the acoustic cue of word-initial phoneme duration and the perceptual effect remains inferential. In addition, other cues to word boundaries were not examined.

The main aim of the present study was thus to explore the degree to which listeners use various acoustic cues to word boundaries in segmentation of continuous speech. To this end, ambiguous Dutch phrases were constructed, containing either a stop-initial word (e.g., the word *pot* in *ze heeft wel eens pot gezegd*, "she said once jar") or a cluster-initial word that matched the stop-initial word together with the preceding [s] phoneme (e.g., the word *spot* in *ze heeft wel een spot gezegd*, "she did say one mockery"). The sentences were, thus, phonemically identical but differed in their precise acoustic-phonetic realization. The sentences were manipulated by cross-splicing, such that the initial stop of the target word and the preceding phoneme [s] (e.g., the [s] and the [p] in *ze heeft wel eens pot gezegd*) were either replaced by a cluster from the cluster-initial word, or by an initial stop and preceding [s] from another recording of the sentence. Acoustic measurements of the ambiguous sequences (e.g., the [sp] in *eens pot* vs. *een spot*) were performed to assess the differences between them. The degree to which a stop-initial word in this context can be discriminated from a cluster-initial word should depend on the acoustic correlates of word boundaries. The eye-tracking paradigm was used to evaluate listeners' ability to distinguish between the two ambiguous sentences. In the eye-tracking paradigm, participants generally hear a sentence and are then shown four objects presented as pictures on a computer screen. Their task is to click on and move the object referred to in the sentence with the computer mouse. Of primary interest was whether participants' fixations to the target picture would differ across the splicing conditions. Subsequently, the acoustic information which participants might be using was determined by correlating their performance in the eye-tracking task with the differences found in the acoustic analyses.

2. Method

2.1. Participants

Twenty-four members of the Max Planck Institute subject panel, native speakers of Dutch, were paid to take part.

2.2. Materials

Twenty stop-initial Dutch nouns referring to picturable objects (e.g., *pot*) were selected, such that the addition of an initial [s] to each word would result in an existing Dutch noun. For example, the addition of an [s] to the Dutch word *pot* results in the word *spot*. Note that the cluster-initial counterpart words were not necessarily picturable nouns. Each target was paired with a cluster-initial picturable noun (the competitor) which overlapped with the first two phonemes of the target's cluster-initial counterpart. For example, for the target *pot*, the competitor *spin* (spider) was selected, overlapping with the first two phonemes of *spot*. Two semantically and phonologically unrelated distractors were assigned to each target and competitor pair (e.g., *vuur* [fire] and *kompas* [compass]). Line-drawing pictures associated with the items were selected from various picture databases. In addition to the 20 experimental item sets, 50 filler sets were constructed. Pictures for the filler trials were selected from the same databases as were used for the experimental trials.

For each experimental item, two recording contexts were constructed. In one of the contexts the target word was mentioned and in the other the target's cluster-initial counterpart was mentioned. The recording contexts were constructed such that the sequences containing the target or its counterpart were identical and, therefore, fully ambiguous. For example, the sentence *ze heeft wel eens pot gezegd* is phonemically identical to the sentence *ze heeft wel een spot gezegd*.

All sentences were read aloud in random order by a female speaker of Dutch in a sound-attenuated booth and recorded on a DAT tape (sampling at 48 kHz with 16-bit resolution). Each sentence was recorded at least four times. The sentences were then re-digitized at a sample rate of 16kHz and edited using Xwaves speech-editor software. For each target word, two spliced versions of the sentence were created. The carrier phrase for both versions consisted of the initial portion of the target sentence (up to the [s], e.g., *ze heeft wel een*) taken from the target recording context, and the final portion of that context (e.g., *gezegd*). For one version (hereafter, the identity-spliced version) the stop (e.g., the [p] in *pot*) and the preceding [s] were taken from another token of the target recording context and spliced onto the carrier phrase. In the other version (hereafter, the cross-spliced version) the stop and the preceding [s] originated from the cluster-initial recording context (e.g., the [sp] from *spot*) (see Table 1). The cross-spliced sentences were thus lexically identical to the identity-spliced sentences, but differed in the origin of the [s] and the following stop (i.e., whether this sequence was taken from the target or the cluster-initial recording context). Cross-spliced sentences were constructed for 19 filler items. Three of the fillers items proved to be problematic and had to be excluded from the experiment. All splicing points were taken at zero-crossings and the splicing manipulation was done very carefully so as to prevent any acoustic artifacts, such as clicks or other distortions.

Table 1: *Stimulus example of the conditions in the experiment.*

Origin of Recording	Spliced version
Identity-spliced condition	
1a. Ze heeft wel eens pot gezegd	Ze heeft wel een**s** **p**ot gezegd
1b. **Ze heeft wel eens pot gezegd**	
Cross-spliced condition	
1a. Ze heeft wel eens pot gezegd	Ze heeft wel een**s** **p**ot gezegd
2. **Ze heeft wel een spot gezegd**	

2.3. Acoustic analyses

Acoustic measurements of the spliced portions of the stimuli (e.g., the [sp] in *eens pot* or in *een spot*) were carried out to evaluate the extent to which their acoustic realization was influenced by the intended meaning. The following measurements were taken: [s] duration, stop closure duration, and Voice Onset Time. Duration measurements were taken from the spectrograms and the waveforms combined, using Xwaves. In addition, RMS energy and Spectral Centre of Gravity (SCG) were measured for the [s] and for the stops. RMS energy was calculated by taking the logarithm of the mean sum of squares of all sample points in the segment. SCG was measured using the built-in function in Praat speech-editor software, which calculates the average frequency from an FFT spectrum over a frequency range from 0 to 10000 Hz. To measure the SCG of [s], the segment was divided into 15 ms intervals, an FFT spectrum was made for each interval (filtering out the frequency range below 1000 Hz to remove any spurious low frequency components) and the SCG of each interval was taken. The maximal SCG was taken as the SCG for the segment.

2.4. Procedure and design

Participants were tested individually. They were first familiarized with the 268 pictures. The pictures appeared on a computer screen, one at a time, along with their printed name, and participants pressed a response button to proceed to the next picture. After this part of the experiment, the eye-tracking system was set up.

Participants were seated at a comfortable distance from the computer screen. Eye movements were monitored using a SMI EyeLink head-mounted eye-tracking system, sampling at 250 Hz. Pictures were presented on a ViewSonic 17PS screen, and the auditory stimuli were presented over headphones using NESU software. Both eyes were monitored, but only the data from the right eye were analyzed.

The structure of each trial was as follows. First, a central fixation dot appeared on the screen for 500 ms. After that, a spoken sentence was presented to the participants and simultaneously a 5x5 grid with pictures appeared on the screen (see Figure 1). Prior to the experiment, participants received written instructions to move the object mentioned in the spoken sentence above or below the geometrical shape adjacent to it, using the computer mouse. The positions of the pictures were randomized across four fixed positions of the grid while the geometric shapes appeared in fixed positions on every trial. Once the picture had been moved, the experimenter pressed a button to initiate the next trial.

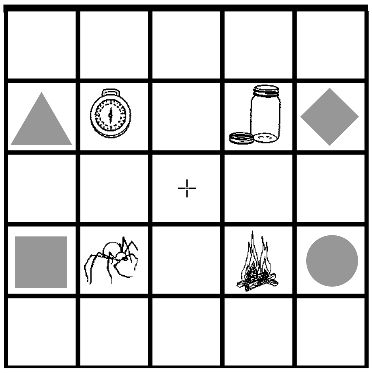

Figure 1: *Example of stimulus display presented to participants.*

Two lists were created, containing the filler and the experimental items. The lists varied on which of the two sentences (i.e., the identity-spliced or the cross-spliced sentence) was presented for each of the experimental items. Within each list, 10 experimental items were assigned to each condition. Twelve random orders were created for the lists, with the constraints that there was always at least one filler item between two experimental items. Participants were randomly assigned to one list.

3. Results

Graphical software was used to display the locations of the participants' fixations as dots superimposed on the four line drawings for each trial and each participant. The timing of the fixations was established relative to the onset of the target word. Fixations on the line drawings were coded as pertaining to the target object, the competitor, or one of the two unrelated distractors, or to anywhere else on the screen. For each trial, fixations were coded from the onset of the target word until the subject had clicked with the mouse cursor on the target picture. Four trials had to be removed from the analysis, because participants clicked on an object other than the target object. The proportions of the fixations were analyzed in 10 ms slices to provide fine-grained information about the time course of lexical activation as the speech unfolded.

Figure 2 presents the proportions of fixations to the target, competitor and distractor pictures, averaged over participants, in the identity-spliced condition and the cross-spliced condition. Fixation proportions to the two unrelated distractors were averaged. Fixation proportions are shown from the splice point (the onset of the [s] preceding the target word) to 1200 ms thereafter.

As is apparent from Figure 2, starting around 350 ms, fixation proportions in the identity-spliced condition rose faster and remained higher than those in the cross-spliced condition. The difference between conditions was statistically tested over a time window extending from 350 to 1200 ms. Over this time interval, the average fixation proportion to the target picture was 61% in the identity-spliced condition and 53% in the cross-spliced condition (F_1 (1,23) = 17.32, $p < .001$; F_2 (1,19) = 7.78, $p < .05$). This demonstrates that the spliced sequences (i.e., the [s] and the following stop) contained fine-grained differences, modulating listeners' lexical interpretation.

To examine these fine-grained differences, acoustic analyses were conducted on the spliced portions in both versions. The results of the acoustic measurements are displayed in Table 2. The results of one-way analyses of variance (ANOVAs) performed on these data are presented in the same table.

The analyses revealed significant differences between the two versions on the following measures: (a) the duration of the [s] in the Target context was significantly shorter than that in the Cluster-initial context; (b) Closure duration was longer in the Target context than in the Cluster-initial context; (c) RMS energy of [s] in the Target context was lower than in the Cluster-initial context and (d) RMS energy of the [t] in the target words was lower than in the cluster-initial words (see Table 2).

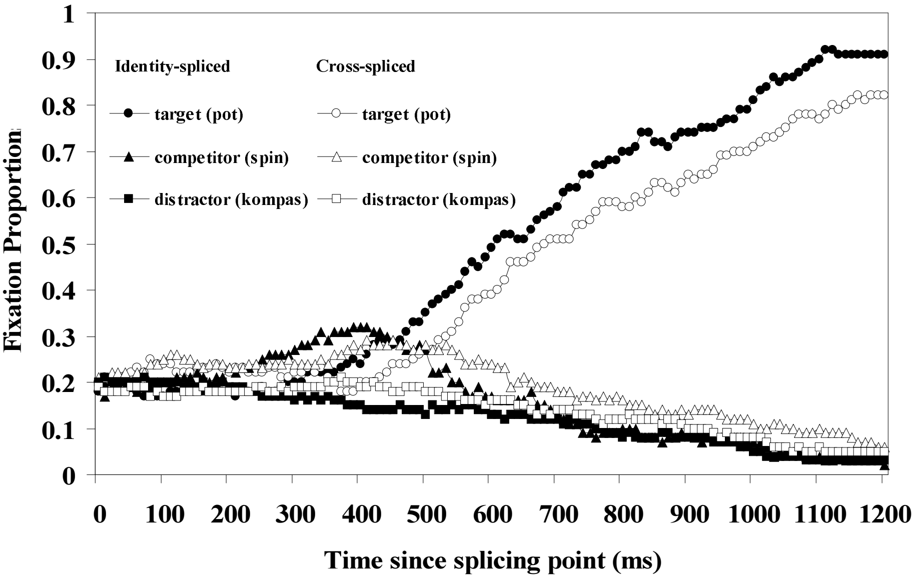

Figure 2: *Fixation proportions over time for identity-spliced and cross-spliced targets, averaged over participants.*

Table 2: *Mean segmental duration (ms), RMS energy (dB), spectral centre of gravity (Hz) and, in brackets, standard deviations of the spliced portions of the experimental stimuli.*

	Target Context	Cluster Context	ANOVA
	eens pot	*een spot*	
Duration			
[s]	87	108	F(1,19) = 24.35
	(12)	(14)	p < 0.001
Closure	90	59	F(1,19) = 41.20
	(23)	(22)	p < 0.001
Voice Onset	23	21	F(1,19) = 1.31
Time	(7)	(7)	n.s.
RMS Energy			
[s]	5.08	5.21	F(1,19) = 10.28
	(0.13)	(0.11)	p < 0.01
stop	4.57	4.70	F(1,19) = 6.97
	(0.13)	(0.16)	p < 0.05
SCG			
[s]	5328	5458	F(1,19) = 1.33
	(316)	(392)	n.s.
stop	1572	1409	F(1,19) < 1
	(1252)	(1159)	n.s.

The acoustic differences between the Target and Cluster-initial sequences are effective cues only to the extent that listeners can perceive these differences and use them in segmentation and word activation. Therefore, for the acoustic measurements for which a significant difference was found, the difference in the measurements for each item was correlated with the perceptual effect for that item (i.e., the difference in the average fixation proportions to the item between the identity-spliced and the cross-spliced conditions in the time window extending from 350 to 1200). These correlations are displayed in Table 3.

Table 3: *Correlation of the difference in the acoustic measurements with the perceptual effect.*

Measurement	Correlation with perceptual effect
Duration of [s]	*r(20)* = .454*
Closure duration	*r(20)* = .285
RMS energy of [s]	*r(20)* = -.033
RMS energy of stop	*r(20)* = -.190

*NOTE: * = p < 0.05*

The correlational analysis showed that out of all the measurements for which a significant difference was found between the two versions, only the duration of the [s] significantly correlated with the perceptual effect. Thus, the data suggest that listeners were using the duration of the [s] as a word boundary cue.

4. Discussion

The main finding of this experiment is that listeners can use phoneme duration as a signal to the location of word boundaries. In the materials used in the experiment, the duration of the [s] was biasing listeners' lexical interpretation of the ambiguous sequence. Other measurements that differentiated the acoustic-phonetic realization of the ambiguous sequences did not correlate with the perceptual data. This suggests that although more cues were available, listeners were attending only to the duration of the [s]. It is however also possible that it is not duration per se that is used by the speech recognition system, but rather some other factor which has not been measured and is highly correlated with duration.

In another eye-tracking experiment [10], the influence of phoneme duration on segmentation has been demonstrated by manipulating the duration of the [s] in the same ambiguous sentences as were used here. The results showed that listeners were slower to identify the stop-initial target object when the duration of the [s] in the spoken signal was lengthened.

These findings add to a growing body of research (e.g., [7,8,9]) showing that fine-grained information in the speech signal can modulate lexical activation. This poses a challenge to current models of spoken-word recognition, none of which take into account the contribution of this type of information.

5. Acknowledgements

I would like to thank Loulou Edelman for lending her voice for the recordings.

6. References

[1] McClelland, J. L., and Elman, J. L. "The TRACE model of speech perception", *Cognitive Psychology* 10: 1-86, 1986.

[2] Norris, D. G. "Shortlist: A connectionist model of continuous speech recognition", *Cognition* 52: 189-234, 1994.

[3] Gaskell, M. G., and Marslen-Wilson, W. D. "Integrating form and meaning: A distributed model of speech perception", *Language and Cognitive Processes* 12: 613-656, 1997.

[4] Klatt, D. H. "The duration of [s] in English words", *Journal of Speech and Hearing Research* 17: 51-63, 1974.

[5] Oller, D. K. "The effect of position in utterance on speech segment duration in English", *Journal of the Acoustical Society of America* 54: 1235-1247, 1973.

[6] Umeda, N. "Consonant duration in American English", *Journal of the Acoustical Society of America:* 61, 846-858, 1977.

[7] Quené, H. "Durational cues for word segmentation in Dutch", *Journal of Phonetics* 20: 331-350, 1992.

[8] Spinelli, E., McQueen, J. M., and Cutler, A. "Processing resyllabified words in French", *Journal of Memory and Language* 48: 233-254, 2003.

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

[10] Shatzman, K. B. and McQueen, J. M. "Durational cues to word boundaries in spoken word recognition", *submitted.*