White, L., Wiget, L., Rauch, O., & Mattys, S.L. (2010). Segmentation cues in spontaneous and read speech. In *Proceedings of the 5th Conference on Speech Prosody, 2010, Chicago* (pp. 100218:1-4).

Segmentation Cues in Spontaneous and Read Speech

Laurence White, Lukas Wiget, Olesya Rauch, Sven L. Mattys

Department of Experimental Psychology, University of Bristol, U.K.

laurence.white@bristol.ac.uk, l.wiget@bristol.ac.uk,
psxop@bris.ac.uk, sven.mattys@bristol.ac.uk

Abstract

Segmentation research asks how listeners locate word boundaries in the ongoing speech stream. Previous work has identified multiple cues (lexical, segmental, prosodic) which affect perception of boundary placement, but such studies have almost exclusively used careful read speech, rather than speech elicited in a natural communicative context. We report development of a segmentation-oriented corpus of spontaneous speech and assess, by comparison with a parallel read speech corpus, how cues such as lexical stress and word-initial lengthening are modulated by the nature of the communicative context, finding evidence in spontaneous speech of contextually-conditioned hypoarticulation that may impact on boundary perception.

Index Terms: speech segmentation, spontaneous speech, rhythm, word-initial lengthening

1. Introduction

Numerous linguistic cues to word boundaries have been identified. Sub-lexical segmentation cues include lexical stress [1], word-initial lengthening [2], glottalisation [3] and phonotactic transition probabilities [4]. Lexical segmentation mechanisms arise from competition between word candidates compatible with various sections of the speech stream and from inferences based on semantic and syntactic expectations [5]. Not all segmentation cues are exploited by listeners at all times, however. In optimal listening conditions, listeners rely on lexical identity and syntactic/semantic structure, and pay less attention to sub-lexical cues [6]. Where lexical and contextual information is unhelpful, sub-lexical cues become relatively more important, with segmental/acoustic cues such as phonotactics and initial lengthening dominant over stress, which, in English, seems to be a last-resort cue when other sources of information are compromised [6].

The occurrence and interpretation of segmentation cues has largely been investigated using carefully articulated read speech. However, elicitation of read speech in the laboratory generally neglects one of the most fundamental aspects of natural conversational speech, the fact that it is goal-directed and interactive. Conversational speech tends to be highly contextualized, with the production and interpretation of utterances being dependent on a quasi-mutual understanding of the foregoing interaction.

The production of spontaneous speech is affected by its interactive, contextualized nature at both segmental and suprasegmental levels. In particular, the degree of articulatory effort in a speaker's utterances – hyperarticulation vs hypoarticulation – has been held to vary as a function of communicative and situational demands [7]. For example, articulatory precision is reduced when contextual information is available: e.g., *nine* is less clearly articulated in *A stitch in time saves nine* than in *The number you are about to hear is nine* [8]. Similarly, stress contrast is attenuated in predictable words compared with unpredictable words [9].

Such findings suggest consequences for the production and interpretation of speech segmentation cues. In particular, cues that are highly salient due to hyperarticulation in non-contextualised speech may be reduced or absent where lexical content is predictable. This predictably could arise as a result of expectation derived from the structure and meaning of the foregoing utterance, or, more straightforwardly, as a result of repetition of words or phrases. Speakers' awareness about the nature of the communicative task must also be considered. Phonetic studies of timing effects using explicitly contrastive speaking tasks – e.g. a study of initial lengthening requiring disambiguation of algebraic bracketings [10] – may overstate their strength in natural speech.

We report development of a set of speech corpora designed to examine the production of segmentation cues in natural conversational speech. Parallel corpora of English spontaneous and read speech allow us to: (1) compare the realisation of word-boundary relevant information in the two speech styles; and (2) test listeners' utilisation of the segmentation cues present in spontaneous speech. We then present phonetic data from our corpora regarding the degree to which word-initial lengthening and stressed syllable lengthening are modulated by the nature of the elicited speech (spontaneous vs read). We also present results of a study examining listeners' perceptions of lexically ambiguous word pairs (e.g. *great anchor* vs *grey tanker*) extracted from read and spontaneous speech, in which the role of repetition in modulating articulatory contrast is explored.

2. Corpus development

2.1. Speakers

Ten speakers of standard Southern British English were recorded. For the conversational speech corpus, speakers participated in pairs; for the read speech corpus they attended individual recording sessions.

2.2. Corpora

2.2.1. Corpus 1: Segmentation-oriented corpus of conversational speech

To elicit spontaneous speech whilst controlling boundary relevant properties, we adapted the Edinburgh Map Task methodology [11], in which two speakers interact conversationally regarding a route around landmarks on a map. Landmark names were one-word or two-word phrases, all paired with similar phrases with which they contrasted in terms of potential word boundary cues (e.g. Figure 1). To avoid speakers having to read landmark names off the map, they were familiarised with the landmarks in an initial training phase until they could reliably name all 108 landmarks without text. The four principal landmark conditions were:

- Cross-boundary allophony (Figure 1): 8 near-homophonous ambiguous two-word phrase pairs (e.g., *great anchor* vs *grey tanker*) and 8 matched

non-ambiguous phrase pairs (e.g., *bright anchor* vs *dry tanker*).

- Presence/absence of a word boundary: 8 pairs of phrases contrasting in the presence or absence of a word boundary (e.g., *seal chair* vs. *wheelchair*).
- Stress vs phonotactics: 16 pairs of two-word phrases contrasting in the stress pattern of the second word: 8 phrases had a strong-weak second word, the predominant English stress pattern, which may favour segmentation (e.g., *cream rickshaw*), whereas the other 8 phrases had a weak-strong second word (e.g., *cream recluse*). Cross-boundary diphone frequencies (calculated from pre-existing corpora) were varied within both stress categories: e.g. /mr/ in *cream rickshaw* has low within-word frequency, favouring segmentation; /br/ in *drab rickshaw* has high within-word frequency, disfavouring segmentation.
- Semantic predictability: 8 pairs of two-word phrases contrasting in the degree to which the first word is associated with a second word common to the two phrases (e.g. *tanker* in *seal tanker* vs *oil tanker*).

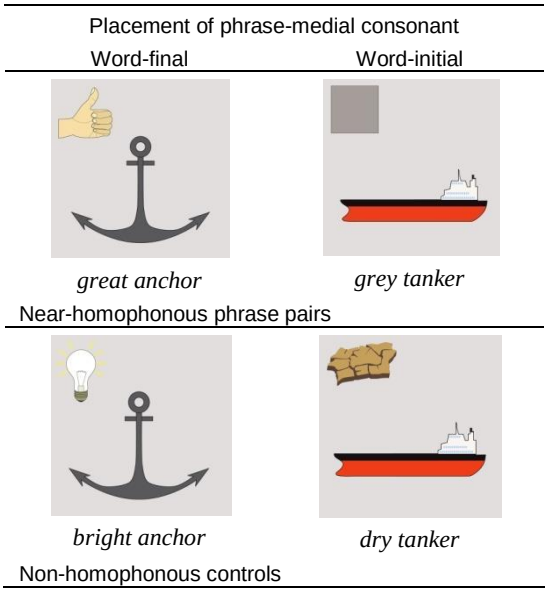

Figure 1: *Example landmarks used in the map description task for the segmentation-oriented corpus of spontaneous speech. Most landmarks, apart from the small number of single-word landmarks, comprised two words. Participants first learned the individual symbols (e.g.* great, grey, anchor, tanker, *etc.) and then learned pairwise combinations.*

Both speakers had maps, but only one speaker's map indicated a route, which s/he had to describe to the other speaker, who was allowed to ask questions and respond to information in a natural conversational manner. There were sixteen maps in total. One speaker acted as describer for eight maps and the roles were reversed for the other eight maps. To avoid explicit disambiguation, contrasting members of phrase pairs (e.g. *great anchor* vs *grey tanker*) were never presented on the same map.

2.2.2. Corpus 2: Parallel corpus of read speech

To allow comparison of the realisation of cues between spontaneous and read speech, all map description utterances containing landmarks were orthographically transcribed and a subset re-recorded as read speech. For the read corpus recordings, we only used those utterances containing the first and second spoken instances of each landmark by each speaker, with minor amendments to the wording of utterance onsets made if required to obtain full self-contained sentences. Each speaker recorded their own landmark utterances in a separate session, after at least a one-month delay. Utterances were presented one at a time on a computer monitor, and speakers were asked to read them in their natural voice at a normal rate. Presentation was self-paced, with corrections to misread sentences prompted either by the experimenter or by speakers themselves.

3. Corpus analyses

A wide range of phonetic analyses of the corpora are in progress. Here we report results relating to two segmentation-relevant durational phenomena discussed in the introduction: contrastive rhythm (i.e. the lengthening of stressed syllables relative to unstressed syllables) and word-initial lengthening.

3.1. Contrastive rhythm

3.1.1. Method

We used two metrics, VarcoV and %V [12], to estimate the degree of stress-related lengthening in spontaneous speech compared with read speech. Eight parallel utterances from the two corpora were selected for each of six speakers. Each utterance was segmented using Praat (http://www.praat.org) into vocalic and consonantal intervals applying standard criteria [12]. Where initial parts of utterances had been reworded for the read session, these were omitted from measurement to keep map and read utterances equivalent. For each utterance, VarcoV was calculated as the standard deviation of vocalic interval duration divided by the mean, and %V as the total proportion of utterance duration made up of vocalic rather than consonantal intervals.

3.1.2. Results

It is first worth noting that articulation rate was consistent between speaking styles: mean articulation rate (i.e. speech rate excluding pauses) was 5.4 syllables per second in spontaneous speech and 5.5 in read speech. A repeated-measures ANOVA with speaker as a between-items factors showed no effect of speaking style [$F(1.42) = 1.02$, n.s.], an effect of speaker [$F(5,42) = 3.64$, $p < .01$] and an interaction between speaker and style [$F(5,42) = 2.60$, $p < .05$]. Unsurprisingly, speakers differed in articulation rate and in how consistent they were at maintaining a constant rate between reading and talking, but overall, read sentences were produced with a natural articulation rate comparable to that used when originally producing the utterances in spontaneous map description.

Within this picture of generally consistent rate between speaking styles, the contrastive rhythm analyses indicated differences between styles in the relative timing of segments (Figure 2). There was a tendency for variation in vocalic interval duration to be greater in spontaneous speech (VarcoV: 62.4) compared with read speech (VarcoV: 59.1) [$F(1.42) = 3.48$, $p = .069$]. There was, however, no effect of speaker on VarcoV scores [$F(5,42) < 1$] and no interaction between speaker and style [$F(5,42) = 1.45$, n.s.].

The vocalic proportion of utterances was significantly greater in map speech (%V: 43.9) than in read speech (%V: 42.4) [F (1,42) = 7.92, p < .01]. There was also an effect of speaker [F(5,42) = 4.69, p < .005] and a trend towards a significant interaction between speaker and style [F(5,42) = 2.09, p = .085]. As shown in Figure 2, speakers varied greatly in the degree to which speaking style affected %V.

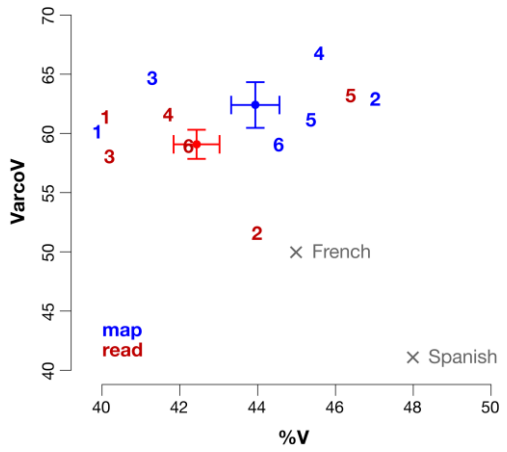

Figure 2: *Contrastive rhythm scores for read speech and spontaneous (map) speech. Mean scores are indicated by the error-barred symbols; individual speakers by numbers. French and Spanish means shown for comparison [12].*

The contrastive rhythm analysis indicates that spontaneous speech is somewhat more vocalic than read speech, and that variation in vocalic interval duration is greater in spontaneous speech. The combination of these two findings suggests that prosodic lengthening processes with extended rather than localised scope (i.e. lengthening of stressed syllables, lengthening of accented words, phrase-final lengthening of stressed and unstressed syllables) may be somewhat exaggerated in spontaneous speech and that this affects vowels more than consonants. Given that phrase-final lengthening tends to be progressive (i.e. greater near the boundary) and that most English syllables end with consonants rather than vowels, amplification of final lengthening would not be expected to affect %V positively. Thus, the most likely sources of this enhanced durational contrast in spontaneous speech are domain-head effects: stressed-syllable lengthening and accentual lengthening. This hypothesis and its consequences for speech segmentation are being investigated in ongoing corpus analyses and perceptual experiments utilising stimuli from Corpus 1 and Corpus 2.

3.2. Word-initial lengthening

3.2.1. Method

For each landmark from the cross-boundary allophony condition, we selected the first fluent versions from the map and from the read corpora uttered by each of nine speakers. Tokens were judged disfluent if there was pause, filled or unfilled, between the two words of the landmarks, or if the landmark was otherwise mispronounced. Thus, in a small number of cases from the map corpus, the second or third version of the landmark was used. In addition, for consistency, only phrase-final versions of landmarks were used. The great majority of landmarks occurred in phrase-final position (i.e. are followed by a perceptible pause), but in a few cases, the first version of the landmark was phrase-medial and these were replaced by subsequent phrase-final tokens.

The duration of the phrase-medial consonant (e.g. [t] in *great anchor*/*grey tanker*) was measured in Praat by inspection of the waveform and spectrogram.

3.2.2. Results

Figure 3 shows the mean duration of phrase-medial consonants in ambiguous (i.e. near-homophonous) phrase pairs (e.g. *grey tanker* vs *great anchor)* and in non-homophonous controls (e.g. *bright anchor* vs *dry tanker*). In the ambiguous condition, a mixed-effects ANOVA showed an effect of position [F(1,262) = 36.52, p < .001], with consonants longer word-initially than word finally. There was also an effect of speaking style [F(1,262) = 4.91, p < .05], and an interaction between speaking style and word position [F(1,262) = 5.45, p < .05]. As shown in Figure 3 (top panel), word-final consonants were shorter in read speech than in map speech, meaning that the durational difference between word-initial and word-final consonants was greater in read speech.

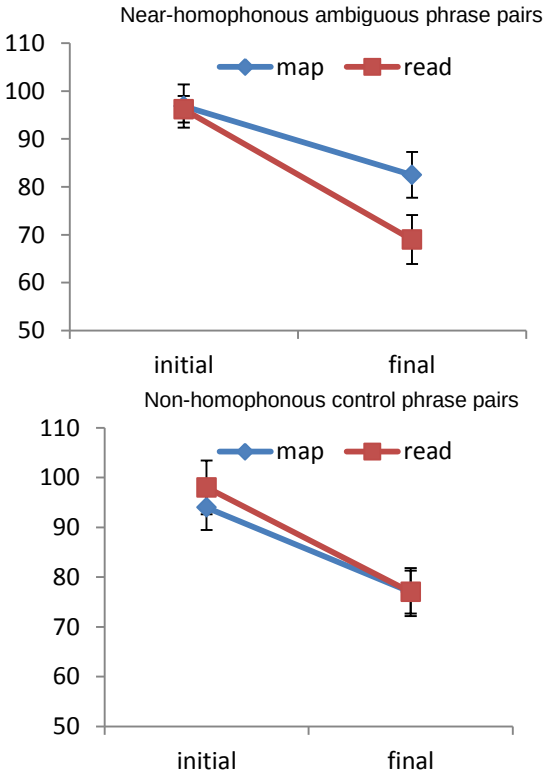

Figure 3: *Consonant duration according to position (word-initial vs word-final), speaking style (spontaneous map dialogues vs read sentences) and phrase type (near-homophonous vs non-homophonous phrases).*

The effect of word position on consonant duration was also evident in the unambiguous phrases (Figure 3), with again a main effect of word position [F(1,262) = 20.79, p < .001]. Here, however, there was no effect of speaking style [F(1,262) = 2.62, n.s.] and no interaction between position and speaking style [F(1,262) < 1]. Thus, when reading, speakers exaggerated the contrast between initial and final consonant duration for ambiguous phrases, but not for comparable unambiguous phrases. Thus, it seems that speakers provide a stronger consonantal cue to word boundary location when phrases are ambiguous, but only in read speech. Studies using explicit contrast in read speech may therefore over-estimate the degree of word-initial lengthening found in spontaneous speech. In both styles and for both phrase types, however,

word-initial lengthening is strongly supported as a potential word boundary cue for listeners. In the next section, we consider the impact of this and other cues – modulated by speech elicitation style (spontaneous vs read) and by repetition – on listeners' interpretation of ambiguous phrases.

4. Perceptual experiment

4.1. Method

We extracted all tokens of near-homophonous phrases from the cross-boundary allophony condition, excluding those with perceptible pauses between words. Fifteen native English speakers listened to all tokens, in random order, indicating on a 9-point scale how ambiguous each token sounded: e.g. "1" was definitely *great anchor*, "9" was definitely *grey tanker*, with a rating of "5" for maximally ambiguous tokens.

4.2. Results

We first considered if the overall pattern of listeners' responses to ambiguous tokens varied between read and spontaneous speech. Wilcoxon signed-ranks test found no statistical difference for phrases with word-final medial consonants (e.g. *great anchor*, mean ratings: map 3.3, read 2.7) and likewise no difference for those with word-initial medial consonants (e.g. *grey tanker*, mean ratings: map 7.8, read 6.8). To test the hypothesis that repetition-induced predictability may tend to induce hypoarticulation, we compared the ambiguity of the first and second realisations of near-homophonous tokens in read and map speech. There was no significant difference between repetitions for read tokens (see Figure 4), but map tokens were judged significantly more ambiguous on repetition than on first utterance: word-final medial consonant phrases (e.g. *great anchor*), $p < .05$; word-initial medial consonant (e.g. *grey tanker)*, $p < .05$.

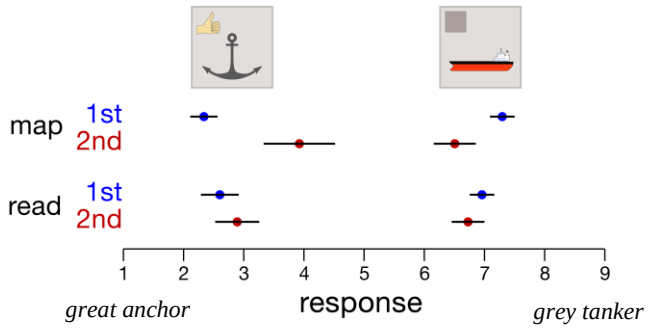

Figure 4: *Listeners' mean ambiguity judgement in response to near-homophonous phrases. Phrases are grouped according to elicitation task (map vs read) and occurrence number (first vs second utterance by each listener). The map symbols illustrate what speakers intended when uttering the phrases scored below. Listeners' responses ranged from "1" (definitely the phrase with a word-final medial consonant) to "9" (definitely the phrase with a word-initial medial consonant).*

5. Conclusions

We have reported the development of a segmentation-oriented corpus of spontaneous speech. A map task allowed generation of interactive dialogues, while the use of landmark phrases contrasting in potential boundary cues allowed us to examine the realisation of such cues in spontaneous speech and to utilise the phrases as stimuli in perceptual studies. Speakers were pre-trained on landmarks to eliminate the need for written text on the maps: recordings revealed that speakers subsequently used landmarks fluently in their map description dialogues. Landmark-carrying utterances were re-recorded as read speech to allow a direct comparison of the realisation of segmentation cues between read and spontaneous elicitation methods. Both spontaneous and read speech corpora will be released for general research use at the end of the project.

The spontaneous and read utterances from the corpora are currently being used in a number of phonetic analyses and perceptual studies of listeners' exploitation of segmentation cues. We reported here on the realisation of two potential timing cues to word boundaries. Word-initial lengthening was robustly observed in both read and spontaneous speech, and whether or not phrases were ambiguous, although the contrast between consonant duration in initial and final positions was particularly marked for ambiguous read tokens, suggesting a degree of relative hypoarticulation in spontaneous speech. This was supported by an experiment examining the perception of ambiguous stimuli, which indicated that tokens in spontaneous, but not read, speech became more ambiguous with repetition. Analyses are in progress to determine the phonetic correlates of this ambiguity. Contrastive rhythm analyses indicated that domain-head lengthening effects may be exaggerated in spontaneous speech, suggesting the possibility that domain-edge cues are attenuated in response to the greater strength of stress cues.

6. Acknowledgements

Research funded by Leverhulme Trust grant F/00182/BG to Sven Mattys. We gratefully acknowledge the support of the Sound to Sense Marie Curie research training network.

7. References

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

[2] Quené, H. (1992). "Durational cues for word segmentation in Dutch." Journal of Phonetics, 20, 331-350.

[3] Nakatani, L.H. & Dukes, K.D. (1977). "Locus of segmental cues to word juncture." Journal of the Acoustical Society of America, 62, 714–719.

[4] McQueen, J.M. (1998). "Segmentation of continuous speech using phonotactics." Journal of Memory and Language, 39, 21-46.

[5] Norris, D., McQueen, J.M. & Cutler, A. (1995). "Competition and segmentation in spoken-word recognition." Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 1209-1228.

[6] Mattys, S.L., White, L., & Melhorn, J.F (2005). "Integration of multiple segmentation cues: A hierarchical framework." Journal of Experimental Psychology: General, 134, 477-500.

[7] Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H&H theory." In W.J. Hardcastle & A. Marchal (Eds.), Speech Production and Speech Modelling (pp. 403-439). Amsterdam: Kluwer.

[8] Lieberman, P. (1963). "Some effects of semantic and grammatical context on the production and perception of speech." Language and Speech, 6, 172-187.

[9] Bock, J. K., & Mazzella, J. R. (1983). "Intonational marking of given and new information: Some consequences for comprehension." Memory and Cognition, 11, 64-76.

[10] Fougeron, C. & Keating, P. A. (1997) "Articulatory strengthening at edges of prosodic domains." Journal of the Acoustical Society of America, 101, 3728-3740.

[11] Anderson, A., Bader, M. et al. (1991). "The HCRC Map Task Corpus." Language and Speech, 34, 351-366.

[12]White, L., & Mattys, S.L. (2007). "Calibrating rhythm: First language and second language studies." Journal of Phonetics, 35, 501-522.