CLEAR SPEECH FACILITATES WORD SEGMENTATION: EVIDENCE FROM EYE-TRACKING

Zhe-chen Guo, Rajka Smiljanic

The University of Texas at Austin
zcadamguo@utexas.edu, rajka@austin.utexas.edu

ABSTRACT

Listener-oriented hyperarticulated clear speech improves word recognition in noise and auditory memory. Using visual-world eye-tracking, this study examined whether clear speech benefits spoken word segmentation and how its effect develops over time. English-speaking listeners heard clear and conversational sentences in which the target word (e.g., *ham*) was temporarily ambiguous with a competitor (e.g., *hamster*) across a word boundary (e.g., *she saw the ham starting*…). Analysis of the listeners' eye fixations indicated that the likelihood of fixating the picture of the target compared to that of the competitor was higher for clear speech than conversational speech. The difference emerged even before segmental information for resolving the ambiguity was available in clear speech. The findings showed that speaking clearly facilitated listeners' discovery of word boundaries. Improved speech segmentation and reduced lexical competition may in part underlie the clear speech processing benefits.

Keywords: Clear speech, speech segmentation, hyperarticulation, eye-tracking

1. INTRODUCTION

Words in fluent speech are not consistently separated by pauses [1]. To successfully recognize words and comprehend the spoken message, the listener needs to find the word boundaries or to achieve speech segmentation. This study aims to provide further insight into such an ability by examining how intelligibility-enhancing hyperarticulated speaking styles affect the listener's speech segmentation.

Past research has revealed a wide variety of signal-dependent and -independent cues used during segmentation (see [2] for a review). The phrase *night rates* can be distinguished from *nitrates* based on allophonic differences [3]. Stressed syllables [4], [5] and vowel lengthening [6], [7] are used to locate word-initial and -final positions, respectively. Signal-independent lexical cues, such as knowing that *anything* is a word, allows listeners to detect the fragment *corri* more rapidly in *anythingcorri* than in *imoshingcorri* [8]. Speech does not even have to be meaningful to be segmented. Exposed to nonsense syllable sequences repeated continuously, listeners can still extract the sequences by tracking syllable co-occurrence frequencies [6], [9].

While these findings improve the understanding of what cues are useful for identifying word boundaries, speech segmentation is rarely explored in connection with phonetic variation in real-word communication (cf. [10]). In daily interactions, speakers dynamically adjust their output along a hypo-hyperspeech continuum, reflecting a balance between speaker- and listener-oriented forces (H&H theory: [11]). Under challenging listening conditions (e.g., noisy environment or non-native listener), they produce intelligibility-enhancing hyperarticulated "clear speech" [12]. Relative to casual or conversational speech, clear speech shows various acoustic-phonetic enhancements (e.g., longer segment duration, vowel space expansion) and benefits perceptual processes including word recognition and auditory memory (for reviews, see [13], [14]).

Recent work has begun to explore these perceptual benefits in finer detail by examining how clear speech impacts listeners' segmentation. In [15]'s artificial language learning study, English listeners heard speech streams containing continuous repetitions of nonsense words and were tasked with segmenting the words based on the statistical patterns of the words' component syllables. Results showed that in quiet, their segmentation was more accurate with the speech streams produced clearly than with those produced conversationally. These findings suggest that the well-established clear speech benefits for word recognition may partly be aided by improved segmentation. Yet, unlike the participants in [15], listeners in real-world communication rarely parse nonsense speech. It remains unclear the extent to which speaking clearly improves segmentation of meaningful words as the speech signal unfolds.

We addressed this issue in a visual-world eye-tracking experiment [16]. Unlike traditional behavioral tasks (e.g., word-spotting [17]), eye-tracking provides rich temporal information of how the evolving speech is integrated to locate word boundaries. In our experiment, listeners heard "lexical garden-path" sentences [18], [19] such as *She saw the ham starting to get crispy and brown*, in which the string *ham st-* overlaps with the unintended *hamster*. Models of spoken word recognition (e.g., [20], [21]) assume that both *ham* and *hamster* are

activated and compete for lexical selection until the signal mismatches *hamster* at the "disambiguation point (DP)" (e.g., onset of /ɑ/ in *starting* in the above sentence). The question of interest concerns whether clear speech provides more word boundary robust cues to allow listeners to consider *ham* over *hamster* and at what point in the evolving speech signal this benefit occurs. To explore this, we tracked English listeners' gaze at related images as they listened to the sentences produced clearly and conversationally.

2. EYE-TRACKING EXPERIMENT

2.1. Hypothesis

Based on [15]'s findings, it was hypothesized that relative to conversational speech, clear speech similarly aids segmentation of meaningful words such that listeners can locate the word boundary more effectively even before the DP is reached. If this was the case, we expected that within the analysis time window (see Section 2.5), the likelihood of fixating the image of the target (e.g., *ham*) as compared with that of fixating the competitor image (e.g., *hamster*) would be significantly higher for clear speech than for conversational speech.

2.2. Stimuli

Critical items were twenty-six pairs of picturable English nouns, each containing a monosyllabic target word (e.g., *ham*) and a disyllabic, initially stressed competitor in which the target was embedded as onset (e.g., *hamster*). The log frequencies per million words of the targets (mean: 2.76) and competitors (mean: 2.06) were matched as closely as possible using the CLEARPOND database [22] to minimize frequency effects [23]. For each pair, a target-bearing sentence was constructed such that the target plus the onset of the next word overlapped partially with the competitor (e.g., *She saw the* ***ham*** *starting to get crispy and brown*). Two additional sentences were created for each pair: one in which the intended word was the competitor (e.g., *She saw the* ***hamster*** *running quickly on the wheel*) and the other in which it was a mono- or di-syllabic distractor word semantically and phonologically unrelated to the target and competitor (e.g., *He wanted to find the* ***turkey*** *under the tree*). The sentences began in a way that would not bias towards either the target, competitor, or distractor. Seventy-eight unique sentences (26 target-competitor pairs × three sentence types) were constructed.

A female native speaker of American English produced the sentences first conversationally and then clearly. For the conversational style, she was instructed to speak "as if to a friend or someone familiar with her voice." For the clear style, she was instructed to speak "as if to someone hard-of-hearing or a non-native speaker." The instructions have been shown to be successful in eliciting the style distinctions [14], [24], as confirmed by a preliminary acoustic analysis showing that, for example, segments in clear speech were longer. The 156 sentence stimuli (78 sentences × two styles) were recorded using a Shure SM10A head-mounted microphone at a 44.1k Hz sampling rate as WAV files.

The stimuli were each paired with a display containing four drawings in the quadrants to form 156 trials. The drawings depicted the target, competitor, a monosyllabic distractor, and a disyllabic distractor, as shown in Fig. 1. In the critical trials, the drawing of the target was mentioned in the sentence while in the filler trials, the mentioned drawing was the competitor or one of the distractors. Drawings of each type appeared in each quadrant roughly equally often. They were taken mostly from normed drawing databases (e.g., [25]) and edited to ensure consistent visual features when necessary.

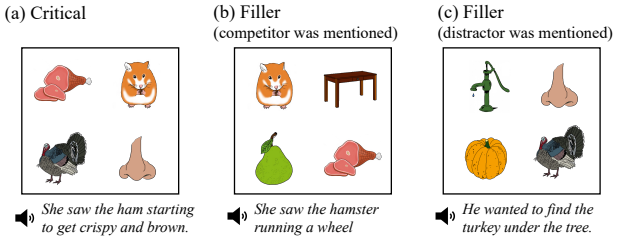

Figure 1: Sample visual arrays of the three trial types

The trials were distributed over two lists of 78 trials such that participants were presented with clear and conversational speech but never with the same sentence in both styles. Each list was divided into three blocks, with block order and trial order within a block randomized. Trials of the three types in Fig. 1 and of the two speech styles occurred (approximately) an equal number of times.

2.3. Procedure

Participants were first familiarized with the drawings and their labels. They were then seated about 60 cm from a computer screen with their heads on a chin rest and assigned to either list. Each trial began with a fixation cross that served as a drift check. When fixated, the cross was displaced by four drawings and three seconds later, a sentence was played in quiet via headphones. Participants clicked the drawing mentioned in the sentence as quickly and accurately as possible. They completed five trials with feedback as practice before the experiment proper. Their right eyes were tracked at 500 Hz with an EyeLink Portable Duo (SR research) eye-tracker, calibrated with a

standard nine-point grid before the practice and each block and when recalibration was needed. The experiment was run using Experiment Builder.

2.4. Participants

Forty-eight native American English speakers participated in the study (29 females; age range: 16-29, mean age: 19.1). All passed a pure-tone screening binaurally at 25 dB HL for 1000, 2000, and 4000 Hz. None reported hearing, speech, or vision disorders. Nine participants were excluded for calibration failure, not fixating any image in almost all trials, or having response accuracy 2.5 standard deviations below the mean for at least one speaking style.

2.5. Data analysis

We tested the hypothesis by comparing proportions of fixations to target versus competitor images in the critical trials within a selected analysis time window. This window began 200 ms after target word onset as it takes about 200 ms to launch a saccade [26]. The end point of the window was sentence-specific but the same for the clear and conversational versions of each sentence—it was the DP of the clear stimulus. This was done as the durations from target onset to the DP in conversational stimuli (mean: 259 ms) could be too short to provide meaningful data when the initial 200 ms was excluded. Thus, since conversational speech was faster, the analysis window included some post-DP segments for the conversational counterpart of each clear stimulus. Such a window provided a stringent and realistic test of the hypothesis: clear speech prior to the DP may contain more robust word boundary cues than conversational speech with post-DP information for ruling out the competitor.

The analysis window of each sentence was divided equally into 10 time bins. Within each bin, proportions of fixations were calculated separately for the target and competitor images and transformed into empirical logits [27]. With this we derived from each trial two empirical-logit curves representing fixation likelihoods over normalized time—one for the target and the other for the competitor.

3. RESULTS

The valid participants (N = 39) clicked with >92% mean accuracy across trial types and styles. Fixations were coded by the eye-tracker's default algorithm as directed to one (or none) of the images on the display and those in the critical trials with correct responses were analyzed as in Section 2.5. Fig. 2 shows the mean empirical logits over the 10 time bins within the analysis window by image type and speaking style.

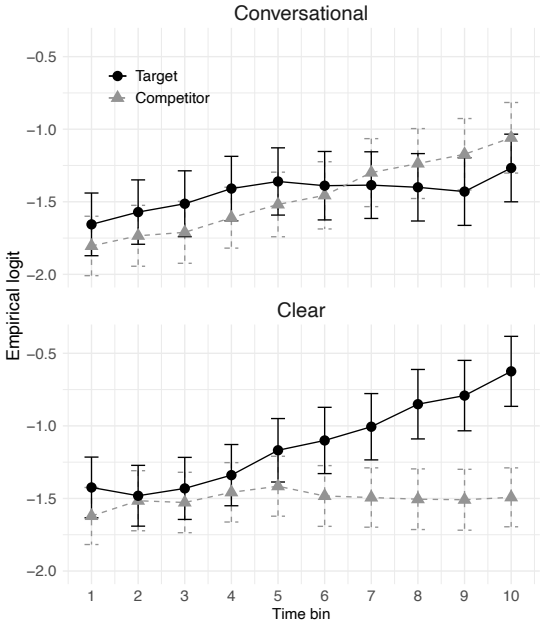

Figure 2: Mean empirical logit at each time bin within the analysis window for fixation data in correct critical trials by speaking style and image type.

Inspecting Fig. 2 revealed that the empirical-logit curve representing the likelihood of target fixations diverged from the curve for competitor fixations to a larger extent for clear speech than for conversational speech. To test the statistical significance of the pattern, we analyzed the curves using generalized additive mixed modeling (GAMM: [28]) implemented in the *mgcv* [29] package of R [30]. Following [31], we modelled the time-varying effects of image type and style and their interaction with smooth functions for three binary predictors: IsTarget (coded as 1 for the curves of the target image and 0 for those of the competitor), IsClear (coded as 1 for the curves of clear speech and 0 for those of conversational speech), and IsTargetClear (coded as 1 for the curves of the target in the clear condition; otherwise, 0). Of interest was the smooth of IsTargetClear, for which positive estimates indicated that target fixation proportions compared with competitor fixation proportions were relatively higher for clear speech than for conversational speech. By-participant and by-sentence factor smooths for the three predictors were included as random effects. Autocorrelated residuals were corrected using an AR(1) error model with $\rho = 0.823$.

All smooth terms were significant ($p < .001$). As these terms were not immediately interpretable, the hypothesis was tested by visualizing the smooth of IsTargetClear over time normalized between 0 and 1. As Fig. 3 shows, its effect was positive and significant from 0.576 to the end. That is, starting around halfway of the analysis window, the difference of the target from the competitor in fixation proportions began to be significantly larger for clear speech than

conversational speech. This result was consistent with the hypothesized clear speech segmentation benefit.

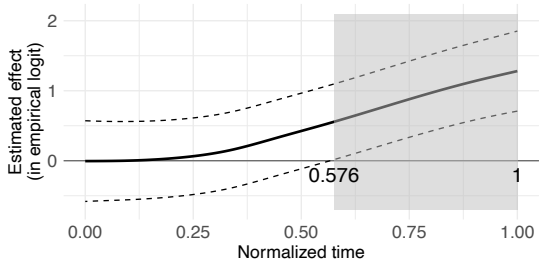

Figure 3: Estimated effect of IsTargetClear over normalized time. The dashed lines mark the 95% confidence interval. The time interval during which the effect was significant is shaded.

4. DISCUSSION

Using visual-world eye-tracking, we examined whether clear speech facilitates word segmentation relative to conversational speech. English-speaking listeners heard clear and conversational sentences with temporary ambiguity between the target and a competitor across a word boundary. Analyzing their eye fixation patterns revealed that the proportions of target fixations as compared with those of competitor fixations were significantly higher for clear speech than conversational speech at 0.576 into the analysis time window. This suggested that clear speech led listeners to locate the word boundary more effectively and consider the intended target word more. Thus, speaking clearly improves not only segmentation of nonsense speech streams as found in [15] but also segmentation of meaningful words during real-time speech processing. Improved word segmentation may in part underlie the clear speech benefits for linguistic and cognitive processes, including word recognition and auditory memory [13], [14].

The findings cannot be attributed to poor intelligibility of conversational speech. This possibility is discounted by a further analysis comparing fixations to the two related images (i.e., target and competitor) versus the unrelated distractor images across the two styles. The results showed that throughout the analysis window, the target and competitor on the one hand were distinguished from the distractors on the other equally well for both styles. This means that the listeners could accurately perceive the critical segments (e.g., *ham st-*) and reject the distractors in conversational speech. The conversational stimuli were just more ambiguous between the target and competitor than the clear ones, resulting in the patterns in Fig. 2.

The current evidence for the clear speech segmentation benefit is noteworthy considering the time window selected for the analysis. As mentioned, the window ended at the DP for the clear sentence stimuli but included some segments after the DP for their conversational counterparts. Word recognition models (e.g., [20], [21]) would posit that during the time window, conversational speech should disambiguate the target from the competitor more than clear speech as post-DP segments mismatching the competitor are available. The results, however, suggested the opposite. As our acoustic analysis found segmental and prosodic enhancements (e.g., longer duration, greater F0 range, etc.) in clear speech, boundary cues like stress [5] and vowel lengthening [19] might also be exaggerated. These enhanced cues then jointly promote activation of the target while reducing competition from unintended competitors—an effect not captured by the lexical competition dynamics assumed in the word recognition models. A pressing goal would be to examine the extent to which each boundary cue contributes to segmentation when enhanced through clear speech modifications.

Our findings further suggest that the relative importance of signal-dependent and signal-independent segmentation cues may be altered under hyperarticulation. It has been proposed that at least in optimal listening conditions, lexical and semantic cues generally outweigh signal-dependent acoustic-phonetic and prosodic cues [10], [32], [33]. Yet, this view is based on experimentation with a single cue from each cue type without consideration of hyperarticulated clear speech. The advantage of clear speech over conversational speech with segmental information for ruling out the competitor implies that listeners may put more weight on the signal-dependent cues enhanced through hyperarticulation, potentially outranking lexico-semantic information. Future work is needed to examine how the clear speech benefit changes in the presence of, for example, semantic context to the target word.

Another further question is to examine the extent to which clear speech facilitates word segmentation in adverse conditions, such as environmental noise. Since noise increases lexical competition from competitor words [34], it is of interest to examine whether the clear speech segmentation benefit will be reduced or delayed in noise as a result. Another issue concerns whether clear speech also aids segmentation for other listener populations such as non-native or hard-of-hearing listeners. The findings would add to the understanding of the mechanism underlying the clear speech intelligibility benefit in realistic communicative contexts.

5. ACKNOWLEDGMENT

The authors would like to thank Madeline Smith, Eliana Spradling, and Madison Rider for their assistance with data collection.

6. REFERENCES

[1] Cole, R. A., Jakimik, J., Cooper, W. E. 1980. Segmenting speech into words. *J. Acoust. Soc. Am.* 67, 1323–1332.

[2] Cutler, A. 2012. *Native Listening: Language Experience and the Recognition of Spoken Words*. The MIT Press.

[3] Gow, D. W., Gordon, P. C. 1995. Lexical and prelexical influences on word segmentation: Evidence from priming. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 344–359.

[4] Cutler, A., Carter, D. M. 1987. The predominance of strong initial syllables in the English vocabulary. *Comput. Speech. Lang.* 2, 133–142.

[5] Tyler, M. D., Cutler, A. 2009. Cross-language differences in cue use for speech segmentation. *J. Acoust. Soc. Am.* 126, 367–376.

[6] Saffran, J. R., Newport, E. L., Aslin, R. N. 1996. Word segmentation: the role of distributional cues. *J. Mem. Lang.* 35, 606–621.

[7] White, L., Benavides-Varela, S., Mády, K. 2020. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *J. Phon.* 81, 100982.

[8] White, L., Melhorn, J. F., Mattys, S. L. 2010. Segmentation by lexical subtraction in Hungarian speakers of second-language English. *Q. J. of Exp. Psycho.* 63, 544–554.

[9] Erickson, L. C., Thiessen, E. D. 2015. Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review* 37, 66–108.

[10] White, L., Mattys, S. L., Wiget, L. 2012. Segmentation cues in conversational speech: robust semantics and fragile phonotactics. *Front. Psychol.* 3, 1–9.

[11] Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W. J., Marchal, A. (eds), *Speech Production and Speech Modelling*. Springer Netherlands, 403–439.

[12] Uchanski, R. M. 2005. Clear speech. In: Pisoni, D., Remez, R. (eds), *The Handbook of Speech Perception*. Blackwell, 207–235.

[13] Smiljanić, R., 2021. Clear speech perception. In: Nygaard, L. C., Pardo, J., Pisoni, D., Remez, R. (eds), *The Handbook of Speech Perception*. Wiley, 177–205.

[14] Pichora-Fuller, M. K., Goy, H., van Lieshout, P. 2010. Effect on speech intelligibility of changes in speech production influenced by instructions and communication environments. *Semin. Hear.* 31, 77–94.

[15] Guo, Z.-C., Smiljanić, R. 2021. Speaking clearly improves speech segmentation by statistical learning under optimal listening conditions. *Laboratory Phonology* 12, 14.

[16] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., Sedivy, J. C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634.

[17] McQueen, J. 1996. Word spotting. *Lang. Cogn. Process.* 11, 695–699.

[18] Davis, M. H., Marslen-Wilson, W. D., Gaskell, M. G. 2002. Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 218–244.

[19] Salverda, A. P., Dahan, D., McQueen, J. M. 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* 90, 51–89.

[20] Marslen-Wilson, W. D. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25, 71–102.

[21] McClelland, J. L., Elman, J. L. 1986. The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86.

[22] Marian, V., Bartolotti, J., Chabal, S., Shook, A. 2012. CLEARPOND: cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS One* 7, e43230.

[23] Dahan, D., Magnuson, J. S., Tanenhaus, M. K. 2001. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cogn. Psychol.* 42, 317–367.

[24] Smiljanić, R., Bradlow, A. R. 2009. Speaking and hearing clearly: talker and listener factors in speaking style changes. *Lang. Linguist. Compass* 3, 236–264.

[25] Duñabeitia, J. A., et al. 2018. MultiPic: a standardized set of 750 drawings with norms for six European languages. *Q. J. of Exp. Psychol.* 71, 808–816.

[26] Fischer, B. 1992. Saccadic reaction time: implications for reading, dyslexia, and visual cognition. In: Rayner, K. (ed), *Eye Movements and Visual Cognition*. Springer, 31–45.

[27] Barr, D. J. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J. Mem. Lang.* 59, 457–474.

[28] Lin, X., Zhang, D. 1999. Inference in generalized additive mixed models by using smoothing splines. *J. R Stat. Soc. Series B Stat. Methodol.* 61, 381–400.

[29] Wood, S. N. 2021. mgcv: mixed GAM computation vehicle with automatic smoothness estimation.

[30] R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing.

[31] van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., Wood, S. N. 2019. Analyzing the time course of pupillometric data. *Trends Hear.* 23, 1–22.

[32] Mattys, S. L., White, L., Melhorn, J. F. 2005. Integration of multiple speech segmentation cues: a hierarchical framework. *J. Exp. Psychol. Gen.* 134, 477–500.

[33] Mattys, S. L., Bortfeld, H. 2017. Speech segmentation. In: Gaskell. M. G., Mirković, J. (eds), *Speech Perception and Spoken Word Recognition*. Routledge, 55–75.

[34] Ben-David, B. M., Chambers, C. G., Daneman, M., Pichora-Fuller, M. K., Reingold, E. M., Schneider, B. A. 2011. Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. *J. Speech Lang. Hear. Res.* 54, 243–262.