SHORTER PHONE DURATION FACILITATES ISOLATED SPOKEN WORD RECOGNITION

Catherine Ford, Filip Nenadić, Daniel Brenner, and Benjamin V. Tucker

University of Alberta
cford1@ualberta.ca, nenadic@ualberta.ca, wobaiden@gmail.com, bvtucker@ualberta.ca

ABSTRACT

Contextually predictable, high frequency, competitor-dense words are often produced with less phonetically contrastive categories in spontaneous speech, often manifested with shorter durations. The present study investigates the role of temporal variation in the recognition of isolated words using the Massive Auditory Lexical Decision (MALD) database. Since additional context is lacking for isolated words, it is hypothesized that processing will be inhibited by either (1) loss of information, i.e., shorter durations of individual phones, or (2) durations that are uncommon for a particular phone (both long and short). A measure of phone temporal variation for each word was calculated and then used to predict response latencies in the MALD dataset. We find, however, that neither hypothesis is supported, as shorter phones are found to facilitate word recognition.

Keywords: word recognition, speech, reduction, auditory lexical decision, lexical processing

Acknowledgements: This project was supported by the Social Sciences and Humanities Research Council of Canada [Insight Grant 435-2014-0678].

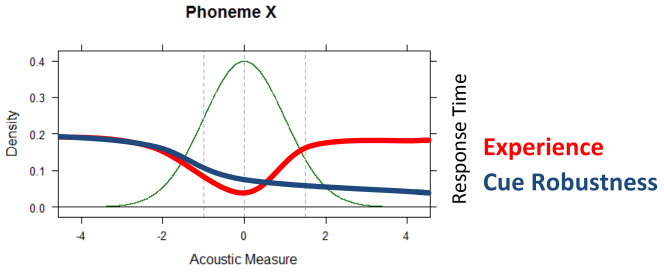
1. INTRODUCTION

Recent investigations of spontaneous speech have shown high frequency and contextually predictable words are often produced with less articulatory effort than careful elicitations [2, 6, 11]. This phenomenon, often referred to as phonetic reduction, has been observed as shorter phone duration, segment deletion, assimilation, and vowel centralization. The present study seeks a quantitative method to measure the degree of reduction found in speech segments and explores how this measure might predict response latency in an auditory lexical decision experiment.

In the investigation of phonetic reduction many studies focus on the duration of the segment – assuming that shorter durations are more reduced and longer durations are less reduced (e.g., [2]). However, in studies of these types there is no standard approach for quantifying the degree of reduction of a segment in relation to productions of the same segment. In other words, it is still an open question as to what the canonical pronunciation of the word is and how we determine the degree of reduction in a word.

The current study focuses on the speech stimuli in the Massive Auditory Lexical Decision (MALD) database [11], and how reduction within the stimuli produced for the task affects spoken word recognition. For the purposes of this investigation, duration is analyzed by calculating the standardized duration of each phone for every phoneme category. The green curve in Figure 1 illustrates a hypothetical distribution of a phoneme X. Based on each phones' standardized duration, a measure of average phone duration can be calculated per word. However, it should be noted that there are many ways in which this variation could be calculated and any number of other acoustic characteristics could be considered.

Figure 1: A hypothetical distribution of "Phoneme X" occurrences duration is given in green, with the duration on the x-axis and frequency density on the y-axis. The secondary y-axis on the right is the hypothetical response time, whereas the red and the blue line represent two hypotheses of how reduction affects response latencies, here referred to as Experience and Cue robustness.



Two competing hypotheses (illustrated in blue and red in Figure 1) were generated based on previous research investigating phonetic variation and lexical processing. First, we could hypothesize that less reduced phones contain more robust cues which facilitate processing (Cue Robustness, e.g., [4, 10]). As a result we would predict that longer phone and word durations would lead to faster response latencies. A second hypothesis would be that the phone durations which are most often observed (e.g., the mean of a distribution) better reflect listener experience (Experience, e.g., [12]). We would then predict that participants will respond more quickly to the average word and phone durations.

2. METHOD

In this paper, we focus on information relevant to the current analysis. The MALD database is described in greater detail in Tucker et al. [11].

2.1 Participants

232 monolingual native Canadian English speakers contributed to the MALD dataset (78% female, age M = 20.11, SD = 2.39). All participants were students at the University of Alberta and received partial course credit for their participation.

2.2 Stimuli

A total of 26,793 word stimuli were organized into 67 separate word sets, and 9,592 pseudoword stimuli were organized into 24 sets. Each word set was matched with 2 pseudoword sets. The total number of experimental lists was therefore 134, each consisting of 400 words and 400 pseudowords. All stimuli were produced by one male speaker of western Canadian English, age 28, who was instructed to read each stimulus as naturally as possible. Pseudoword stimuli were presented to the speaker using IPA transcription with stress represented. Recordings were force aligned using the Penn Forced Aligner [15], allowing for automatic extraction of start and end point of each phone.

2.3 Procedure

The experiment was conducted in sound-attenuated booths with a computer monitor, headphones and a button box. Each participant completed one list that was presented using E-prime experimental software [9].

The task required participants to listen to each stimulus and press the appropriate button with their dominant hand if the stimulus is a real word in English, and to press the opposing button with their non-dominant hand if it is not a real word in English. Responding during a stimulus would result in the interruption of its presentation, leading the experiment to automatically proceed to the following stimulus. If a participant did not respond within 3 seconds of stimulus onset, the experiment would automatically continue to the next stimulus. Stimuli were presented in a random order. Participants could participate in a total of 3 sessions, responding to a different list every time. A total of 284 sessions were recorded resulting in approximately 4 responses per stimulus.

3. ANALYSIS

Phone durations were centered across occurrences within each phone category. Stressed and unstressed vowels were standardized separately, as stress influences vowel duration. Using this standardized duration, we calculated the average phone duration for each MALD word stimulus.

The analysis was performed using generalized additive mixed modelling in *R* [8], using the packages mgcv [14] and itsadug [13]. We first tested a number of baseline models that considered trial number, number of word stimuli in a row including the current stimulus (hereafter "word run length"), number of phones, number of syllables, phonological Levenshtein distance, phonological neighbourhood density, temporal uniqueness point, log-transformed word duration (in ms), and log-transformed frequency increased by 1 from COCA [3] (hereafter "COCA frequency"). Random effects for particular words were not included, as the number of responses per word was low, while the number of words analyzed was too large for models to converge. Random smooths of trial number per participant were included. The dependent variable was response time inverse-transformed to approximate normality (-1000/RT; see [1]). Analysis was conducted on correct responses only. Responses faster than 500 ms and slower than 2500 ms were excluded. The best baseline model was then augmented by the aforementioned mean standardized phone duration.

4. RESULTS

4.1 Correlations

Control predictors mostly correlated lower than $r = |0.3|$. However, there were high correlations between log-transformed duration, temporal uniqueness point, number of phones, number of syllables, and phonological Levenshtein distance. Of the four, only log-transformed duration and temporal uniqueness point were kept in the model. Mean standardized phone duration of a word did not correlate above $r = |0.15|$ with any of the controls predictors.

4.2 Models

The best baseline model included random slopes for trial number per subject, and smoothed effects of trial number, word run length, phonological neighbourhood density, COCA frequency, log-transformed duration, and temporal uniqueness point. Tensor effects of COCA frequency with phonological neighbourhood density and temporal uniqueness point were also included. Results reported in Table 1 indicate that response latencies became shorter as the experiment progressed (Trial), and if given multiple word stimuli in a row (Word Run Length). Participants responded faster to high frequency words (COCA Frequency), and slower to high neighbourhood density words (Phonological ND). Longer words resulted in longer response latencies (log Duration), simply because participants had to wait longer to hear these stimuli (response time was measured from stimulus onset). Temporal uniqueness point (Temporal UP) has an interesting effect on response times, with earlier uniqueness points initially resulting in slower response times. For later uniqueness points, however, response times become faster. Tensor effects indicate that COCA frequency has a stronger effect than either temporal uniqueness point or neighbourhood density, but this is attenuated by these other predictors' values.

Figure 2: Smoothed effect of mean standardized phone duration on response latencies.

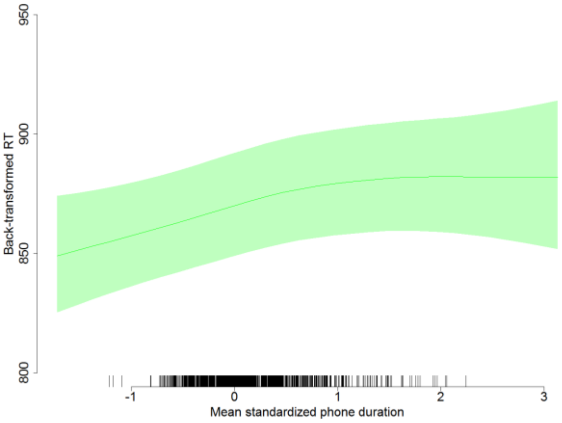


Table 1: The augmented GAMM model containing the predictors from the best baseline model and the mean standardized phone duration of a word. All effects are significant.

Parametric coefficients:			
	Est.	Std. Er.	t value
(Intercept)	-1.13	0.01	-168.6
Approximate significance of smooth terms:			
	edf	Ref.df	F
s(Trial, Subject)	1106.86	2078	12.64
s(Trial)	8.14	8.69	33.11
s(Word Run Length)	2.84	3.54	92.31
s(log Duration)	6.13	7.23	553.38
s(Temporal UP)	6.60	7.61	7.59
s(Phonological ND)	5.26	6.25	30.24
s(log Frequency)	6.32	7.39	337.23
ti(log Frequency, Phonological ND)	8.18	9.57	2.61
ti(log Frequency, Temporal UP)	5.24	6.85	15.59
s(mean standardized phone duration)	2.64	3.39	20.97

Notes: R-sq.(adj) = 0.316; Deviance explained = 32.4%; -REML = -14055; Scale est. = 0.04; n = 101386.

The temporal variation measure developed for this study was also a significant predictor of response latencies. Shorter response latencies were connected with shorter mean standardized phone durations (Figure 2).

5. DISCUSSION

Our results indicate that as temporal variation increases the response latencies decrease. We tested several other variations (calculations) of temporal variation not reported here and they all resulted in a similar outcome. The present results suggest that words with more reduced phone durations facilitate processing of word stimuli in MALD, which does not support either of our working hypotheses. Some results in previous literature have shown that words that are more reduced are responded to more slowly (supporting cue robustness; e.g., [4, 10]). The analysis presented here, however, differs from these previous studies in that reduction was classified based on the duration of each phone in comparison to all other occurrences of that phone across the MALD dataset. A measure of word reduction was then calculated as the mean standardized duration of phones contained in each word.

However, we should also keep in mind that MALD stimuli were recorded in a laboratory setting as a word list. In order to better determine the full variability of phone durations encompassed in MALD, a comparison of MALD productions to corpus data is necessary. Because of potential differences between the MALD stimuli and actual casual speech, it is possible that these results do not contradict previous literature after all. Perhaps the MALD stimuli include fewer extreme reductions, which retain less acoustic information and are more difficult to process. The more reduced productions in MALD contain reductions typical to laboratory speech. In other words, a ‘reduced’ MALD stimulus could be considered only mildly reduced by everyday casual speech standards. As a result, the more typical productions found in the MALD dataset are responded to most quickly, and these have shorter mean standardized phone durations. This interpretation would support the hypothesis that listeners are using their experience to respond to stimuli, and thus more common productions will be processed faster. Therefore, the results captured in the current study may describe responses to a different part of the reduction spectrum than previous studies.

It is also possible that listeners, as they become familiar with the task, recognize that word stimuli are more likely to have reduced phones within them than pseudoword stimuli. Our predictor of reduction may in part capture facilitation in processing as listeners equate shorter durations with “word-likeness”. This, however, should be investigated by looking into pseudoword phone duration and how it relates to participant response latencies to pseudowords [5].

6. CONCLUSION

The current study illustrates that processing of reduced phones is more complicated than researchers initially believed. It is possible that listeners are using statistical information about word productions, and use this stored information to help recognize speech and overcome challenges introduced by variation in pronunciation. Additionally, this study demonstrates that mean standardized phone duration can be used as a measure of reduction within auditory lexical decision tasks, and that even careful speech may have some degree of reduction.

REFERENCES

1. Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12-28.
2. Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.
3. Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159-190.
4. Ernestus, M., Baayen, H., & Schreuder R. (2002). The Recognition of reduced word forms. *Brain and Language, 81(1-3)*, 162-173.
5. Kelley, M., & Tucker, B. V. (2017). *Recognition of spoken pseudowords*. Presentation, Western Conference on Linguistics, USA.
6. Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighbourhood activation model. *Ear and Hearing, 19*(1), 1-36.
7. Plug, L (2011). Phonetic reduction and informational redundancy in self-initiated self-repair in Dutch. *Journal of Phonetics, 39*(3), 289-297.
8. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
9. Schneider, W., Eschman, A., & Zuccolotto. (2012). E-Prime reference guide. Pittsburgh: Psychology Software Tools Inc.
10. Tucker, B. V. (2011). The effect of reduction on the processing of flaps and /g/ in isolated words. *Journal of Phonetics, 39,* 312-318.
11. Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, 1–18.
12. Tucker, B. V., & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon, 11*(3), 375–400.
13. Van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2016). itsadug: Interpreting time series and autocorrelated data using GAMMs. R package version 2.2.
14. Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.
15. Yuan, J. & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics ’08.*