LISTENER PREFERENCE IS FOR REDUCED DETERMINERS THAT ANTICIPATE THE FOLLOWING NOUN

Phil J. Howson & Melissa A. Redford

The University of Oregon
philh@uoregon.edu; redford@uoregon.edu

ABSTRACT

This study examines the effects of determiner reduction and coarticulation on the perceived naturalness of resynthesized *shock-the-geek* (*V-the-N)* sequences. The determiner, equally spaced between monosyllabic *V* and *N*, was manipulated in 3 experiments along a 7-step continuum: (1) duration varied from 0.25x the original duration to 4x this duration; (2) amplitude varied from 55 dB to 85 dB; (3) schwa formants varied from completely overlapped with the vowel in *V* to completely overlapped with the vowel in *N*. Listeners rated *V-the-N* sequences with reduced duration and intensity and more anticipatory coarticulation more favourably than sequences with increased duration and intensity and more preservatory coarticulation. These results are consistent with a listener preference for the production of supralexical chunks that adhere to morphosyntactic rather than metrical structure.

Keywords: coarticulation, reduction, speech perception, grammatical words, speech rhythm

1. INTRODUCTION

The stress-timed rhythm pattern of most English dialects entails grammatical word (e.g. *a*, *the*) reduction. The phonetic correlates of reduction are duration and amplitude [24]. Grammatical words are typically reduced irrespective of other factors [2, 3, 29], but factors such as planning difficulties and prosodic context have been shown to affect reduction [18]. Duration and amplitude measures of grammatical word production indicate that school-aged children do not reduce grammatical words to the same extent as adults [25], which may explain why the acquisition of English speech rhythm is protracted [1, 27]. The current study investigates this possibility further by asking whether grammatical word reduction is in fact important to listeners, since rhythm is fundamentally a perceptual notion. If listeners expect grammatical word reduction, then increasing their duration and intensity should lower listeners' naturalness ratings of the speech, while decreasing their duration and intensity should increase these ratings.

Of course, there is more to rhythm than timing [4, 5, 8]. Prosodic theory argues for the relevance of constituent structure. One important constituent is the prosodic word. Although prosodic words are often equivalent to the lexical word, they are also formed when a grammatical word is cliticized (chunked) with an adjacent content word. When possible, the chunking pattern follows metrical structure. For English, this means that unstressed grammatical words are, by hypothesis, optimally cliticized/chunked with a preceding content word when that content word is monosyllabic [21]. The current study uses listener preferences to investigate this hypothesis further.

In so far as the prosodic word is a production unit [21, 31], then a strong cue to cliticization may be coarticulation [25]. This assumption follows from the well-accepted view that coarticulation, especially anticipatory coarticulation, is largely planned [see 30, 14, 28] and the further assumption that the prosodic word is the relevant planning domain [31, 19]. Thus, in the present study, we manipulated the schwa formant values associated with the definite article, *the*, to create stimuli where this grammatical word was maximally coarticulated with a preceding monosyllabic verb or with a following monosyllabic noun. If listeners expect prosodic words that conform to metrical structure, then they should prefer when *the* is minimally coarticulated with the following noun.

2. METHODS

2.1. Participants

Participants were 150 native speakers of English (*M* = 36 years, *SD* = 10 years), who were recruited using Amazon's mTurk [10]. Sixty-three participants self-identified as female and 87 as male. Participants had no self-reported history of speaking or hearing impairments.

2.2. Stimuli

Stimuli were created by asking four native speakers of a west-coast variety of American English (2 male and 2 female) to produce the phrase *I shock the geek today* without prosodic breaks and at a normal

speaking rate. Elicitations were recorded in a sound attenuated booth using an Audio Technica Lo-Z Condenser and a Tascam US-144 MKII Audio/Midi Interface in Praat [8]. The stimuli used only the *shock the geek* section of the sentence. The overall temporal pattern of this section was controlled by spacing the 100 milliseconds (ms) from the verb and 100 ms from the noun.

Grammatical word (i.e., *the*) duration was manipulated in Praat [8] to create a 7-step within-speaker continuum. Speakers' average duration for the determiner was calculated and used as step 4. For steps 5-7, duration was increased 2, 3, and 4 times, respectively. For steps 1-3, duration was reduced 0.25, 0.33, 0.5 times, respectively.

Grammatical word intensity was also manipulated to create a 7-step within-speaker continuum. The base intensity for the sentence was first normalized to 70 dB. The determiner was then varied from 55 dB to 85 dB at 5 dB intervals.

Finally, the formant values in the schwa vowel of the grammatical word were manipulated to create a 7-step within-speaker continuum from fully coarticulated with the preceding verb to fully coarticulated with the following noun. First, the average F1 to F3 formant values for [ə] in *the*, [ɑ] in *shock*, and [i] in *geek* were calculate for each speaker. Then the average formant measures for the schwa were converted to the nearest half Equivalent Rectangular Bandwidth [ERB; 13] and used as the starting point for the within-speaker continua. Each continuum then had 3 steps above and 3 steps below the average formant measure. Each step was a distance of 1 ERB from the preceding one. The end-points, step 1 and step 7, represented ERB values similar to the average formant values for [ɑ] and [i], respectively. Table 1 and 2 presents the average duration and F1, F2, and F3 for each of the steps.

Table 1: Summary average duration (in *seconds*) and F1, F2, and F3 (in *Hz*) for steps 1-4.

	1	2	3	4
Duration	0.019	0.024	0.037	0.074
F1	750	650	560	480
F2	1256	1425	1612	1822
F3	1690	1908	2151	2421

Table 2: Summary average duration (in *seconds*) and F1, F2, and F3 (in *Hz*) for steps 5-7.

	5	6	7
Duration	0.148	0.222	0.296
F1	407	342	284
F2	2055	2315	2601
F3	2722	3058	3431

2.3. Procedure

Stimuli were blocked by manipulation to create 3 different conditions: a duration manipulation condition, an intensity manipulation condition, and a formant manipulation condition. The design was between-groups, meaning that different groups of 50 participants listened to all the stimuli in each condition. These participants were instructed to wear headphones set to a comfortable listening volume, which they adjusted during a preliminary task in which they were required to input different words that were played to them over the headphones. Participants were then instructed in the main task, which was to rate tokens on a naturalness scale from 1 (least natural) to 7 (most natural). They were then given 7 practice trials to ensure they understood the task. The experiment then began. Stimuli were blocked by speaker. Participants heard each of the four speakers in a randomized order. Stimuli were also randomized within each speaker block. Participants who took part in one condition were not able to take part in another condition.

2.4. Analysis

The rating data were analysed in a linear mixed-effects model using the lme4 package [7] and the R^2 metric was calculated using the MuMIn package [6] in R [23]. A second order polynomial was used for the main effect Step since visual inspection of the data strongly suggested a quadratic relation with this factor. There was a random intercept for participant and random slope for the second order polynomial, Step. The anova() function was use to test for model significance. The lmerTest package [9] was used to estimate the degrees of freedom with Satterthwaite's method [26]. Post-hoc analyses were also performed with Holm [17] correction. Plots were created using ggplot2 [32].

3. RESULTS

The results are presented by condition: duration, intensity, and coarticulation.

3.1. Grammatical word duration

The overall mixed-effects model results indicated a significant effect of the duration manipulation on listener ratings [$F(2, 49) = 324.82$, $p < 0.001$]. Moreover, the model explained a substantial amount of the rating variance [$R^2 = 0.47$]. These overall results are shown in Table 3. Post-hoc analysis revealed that Step 1 (3.71) was not significantly different from Steps 6 (3.80; $p = 0.803$) and 7 (3.62; $p = 0.803$), but Steps 1, 6, and 7, were significantly

lower than Steps 2 (4.90; $p < 0.001$), 3 (4.93; $p < 0.001$), 4 (4.94; $p < 0.001$), and 5 (4.80; $p < 0.001$). Steps 2, 3, 4, and 5 were not found to be significantly different from each other ($p > 0.05$). Table 4 presents the mean ratings by Step, which are also shown in Fig. 1 with the 95% confidence interval.

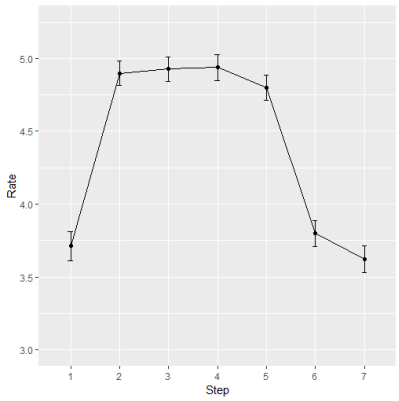
Table 3: Summary of the linear mixed-effects model for the duration manipulation condition.

	Est.	SE	*t-val*	*p-val*
Intercept	4.38	0.14	31.38	< 0.001
Step, 2-1	-17.01	2.92	-5.82	< 0.001
Step, 2-2	-46.37	2.23	-20.71	< 0.001

Table 4: Mean rating and standard deviation (SD). Continuum was short (Step 1) to long (Step 7).

Step	Rating	SD
1	3.71	(1.77)
2	4.90	(1.48)
3	4.93	(1.48)
4	4.94	(1.56)
5	4.80	(1.51)
6	3.80	(1.57)
7	3.62	(1.63)

Figure 1: Line plot for the ratings of each step along the duration continuum. The whiskers show the 95% confidence interval.



3.2. Grammatical word intensity

The overall mixed-effects model results again indicated a significant effect of the manipulation on ratings [$F(2, 95.47) = 12.86$, $p < 0.001$]. Again, the model explained a substantial amount of the rating variance [$R^2 = 0.43$]. These overall results are presented in Table 5. Mean rating data by Step are presented in Table 6 and in Fig. 2. Post-hoc tests confirmed that there were no significant difference between Steps 2 (4.44), 3 (4.53), 4 (4.56), and 5 (4.49, $p > 0.05$), but Steps 1 (3.86), 6 (4.20), and 7 (3.68) were all lower than Steps 2, 3, 4, and 5 ($p < 0.05$). Step 6 was higher than Steps 1 and 7 ($p < 0.01$).

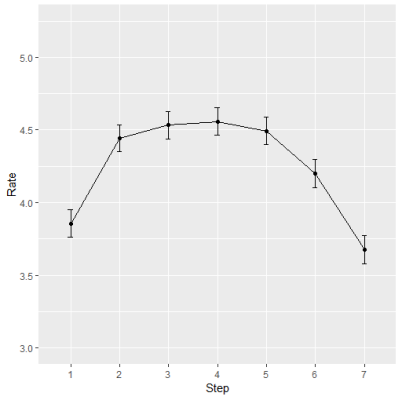
Table 5: Summary of the linear mixed-effects model for the intensity manipulation.

	Est.	SE	*t-val*	*p-val*
Intercept	4.25	0.14	30.42	< 0.001
Step, 2-1	-7.00	1.80	-3.88	< 0.001
Step, 2-2	-28.92	5.75	-5.03	< 0.001

Table 6: Mean rating and standard deviation (SD). Continuum was 55 dB (Step 1) to 85 dB (Step 7).

Step	Rating	SD
1	3.86	(1.69)
2	4.44	(1.65)
3	4.53	(1.66)
4	4.56	(1.65)
5	4.49	(1.67)
6	4.20	(1.67)
7	3.68	(1.72)

Figure 2: Line plot for the ratings of each step along the intensity continuum. The whiskers show 95% confidence intervals.



3.3. Grammatical word coarticulation

The overall mixed-effects model results indicated that the coarticulation manipulation also had a significant effect on listeners' ratings [$F(2, 95.37) = 17.21$, $p < 0.001$]. The model R^2 was 0.35. Table 7 presents the overall results. Mean rating data by Step are presented in Table 8 and in Fig. 3. Post-hoc tests revealed that Step 1 (3.46) was rated lower than every other step ($p < 0.001$), as was Step 2 (3.71), except it was rated higher than Step 1 ($p < 0.001$). Step 3 (4.05) was not significantly different from Steps 4-7 ($p > 0.05$). Steps 4 (4.15) and 5 (4.20) were significantly higher than Step 7 (3.92; $p = 0.005$, $p < 0.001$), but were not significantly different than Step 6 (4.09; $p > 0.05$).

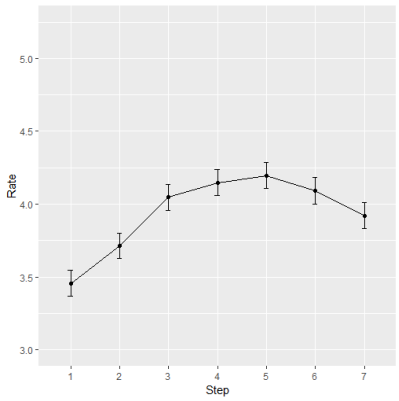
Table 7: Summary of the linear mixed-effects model for the formant manipulations.

	Est.	SE	*t-val*	*p-val*
Intercept	3.94	0.12	31.72	< 0.001
Step, 2-1	15.08	2.72	5.54	< 0.001
Step, 2-2	-16.74	3.00	-5.57	< 0.001

Table 8: Mean rating and standard deviation (SD). Continuum was from [ɑ]-like (Step 1) to [i]-like (Step 7).

Step	Rating	SD
1	3.46	(1.58)
2	3.71	(1.53)
3	4.05	(1.57)
4	4.15	(1.57)
5	4.20	(1.58)
6	4.09	(1.61)
7	3.92	(1.60)

Figure 3: Line plot for the ratings of each step along the formant continuum. The whiskers show 95% confidence intervals.



4. DISCUSSION

The rating results suggest that listeners rate *V-the-N* sequences in which *the* is reduced and coarticulated with the following noun higher than sequences that are not. Specifically, listeners rated longer determiners as less natural than shorter determiners, with the exception of stimuli with the shortest *the*. Listeners also rated less intense determiners as more natural than more intense determiners, with the exception of stimuli with a barely audible *the*. Finally, listeners strongly preferred sequences where formant cues suggested that the determiner was produced with the following noun rather than with the preceding verb.

Whereas the duration and intensity findings conform to expectations based on English speech rhythm patterns, the coarticulation results run counter to these expectations. Rather than suggesting higher naturalness judgement for grammatical word chunking along metrical lines, the data suggest a higher ratings for chunking along morphosyntactic lines. Such an internalized expectation might have a functional explanation: listeners may expect coarticulatory patterns that provide cues to upcoming information. This possibility is consistent with work on children's speech processing. For example, Mahr, McMillan, Saffran, Weismer, & Edwards [22] found that 18- to 24-month-olds looked at the correct image associated with a target noun significantly sooner when the preceding determiner contained coarticulatory cues to the upcoming noun.

Alternatively, listener ratings for sequences in which *the* was more coarticulated with the noun compared to the verb could also reflect an expectation for patterns that closely mimic natural speech. Note that listener ratings were highest in stimuli where the duration and intensity manipulation resulted in a value that was closest to the natural value (i.e., Step 4). Similarly, naturally produced *the* was typically more coarticulated with the following noun than with the preceding verb, which could explain listener preference for stimuli in which the formant values in the determiner were more similar to the following noun than to the preceding verb.

Listeners' higher rating for stimuli that are more similar to natural speech could also signal a processing strategy that references motor patterns. After all, it is by now well-established that speech perception tasks activate motor cortex [11, 12, 16]. The current results might therefore suggest that listeners refer to internalized speech motor plans during speech processing. If this is the case, then we note with interest that listener preference for higher degrees of anticipatory coarticulation than preservatory coarticulation could reflect an acquired plan that is encoded not only for gestures and their targets but how these gestures correctly overlap within continuous speech. The internalized motor plan might also reflect listeners' expectations for how the combinatory nature of different gestures will affect the acoustic signal.

Overall, the present results strongly suggest that prosodically-conditioned grammatical word reduction and cliticization is expected in English. Further research is needed to understand different expectations for child and adult speech to understand the development of reduction and speech planning.

5. ACKNOWLEDGEMENTS

This work was funded by grant number R01HD087452 awarded to Melissa Redford.

6. REFERENCES

[1] Allen, G. & Hawkins, S. 1980. Phonological rhythm: definition and development. In, Yeni-Komshian, G. H., Kavanagh, J. F., & Ferguson, C. A. (eds.), *Child phonology: Production*, pp. 227-256. New York: Academic Press.

[2] Arnon, I. & Snider, N. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62(1), 67-82.

[3] Arnon, I. & Cohen Priva, U. 2013. The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon* 6(2), 302-324.

[4] Arvaniti, A. 2009. Rhythm, timing, and the timing of rhythm. *Phonetica* 66, 46-63.

[5] Arvaniti, A. 2012. The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* 40(3), 351-373.

[6] Barto, K. 2018. *MuMIn* [R package]. v.1.40.4.

[7] Bolker, B. 2018. *lme4* [R package]. v.1.1-19.

[8] Boersma, P. & Weenink, D. 2018. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.43.

[9] Brockhoff, P. B. 2018. *lmerTest* [R package]. v.3.0-1.

[10] Buhrmester, M., Kwang, T., & Gosling, S. D. 2011. Amazon's Mechanical Turk. A new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science* 6, 3-5.

[11] Callan, D. E., Kent, R. D., Guenther, F. H., Vorperian, K. H. 2000. An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research* 43(3), 721-736.

[12] Evans, S. & Davis, M. 2015. Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cerebral Cortex* 25(12), 4772-4788.

[13] Glasberg, B. R. & Moore, B. C. J. 1990. Derivation of auditory filter shapes from notched-noise data. *Hearing Research* 47, 103-138.

[14] Guenther, F. H., Hampson, M., & Johnson, D. 1998. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review* 105(4), 611-633.

[15] Hayes, B. 1982. Extrametricality and English Stress. *Linguistic Inquiry* 13, 227-276.

[16] Hickok, G. & Peoppel, D. 2015. Neural basis of speech perception. In, Celesia, G. G. & Hickok, G. (eds.), *Handbook of clinical neurology vol. 129*, pp. 149-160. Elsevier B. V.

[17] Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2), 65-70.

[18] Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymond, W. 1998. Reduction of English function words in Switchboard. *Proceedings of the International Conference on Spoken Language Processing*, 3111-3114.

[19] Krivokapić, J. 2014. Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences* 369(1658), 20130397.

[21] Levelt, W. J. M., Roelofs, A., & Meyer, A. S. 1999. A theory of lexical access in speech production. *Behavioural and Brain Sciencs* 22(1), 1-38.

[22] Mahr, T., McMillan, B. T. M., Saffran, J. R., Weismer, S. E., & Edwards, J. 2015. Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition* 142, 345-350.

[23] R Core Team. 2018. *R: a language environment for statistical computing*. R Foundation for Statistical Computing. Vienna: Austria.

[24] Ramus, F., Nespor, M., & Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 72, 1-28.

[25] Redford, M. A. 2018. Grammatical word production across metrical cntexts in school-aged children's and adults' speech. *Journal of Speech, Language, and Hearing Research* 61, 1339-1354.

[26] Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2, 110-114.

[27] Sirsa, H. & Redford, M. A. 2013. The effects of native language on Indian English sounds and timing patterns. *Journal of Phonetics* 41, 393-406.

[28] Tilsen, S., Spincemaille, P., Xu, B., Doerschuk, P., Luh, W. M., Feldman, E. 2016. Anticipatory posturing of the vocal tract reveals dissociation of speech movement plans from linguistic units. *PLoS ONE* 11(1), e0146813.

[29] Tremblay, A. & Tucker, B. V. (2011). The effect of N-gram probabilistic measures on the recognition and production of four-word sequences. The mental Lexicon 6, 302-324.

[30] Whalen, D. H. 1990. Coarticulation is largely planned. *Journal of Phonetics* 18, 3-35.

[31] Wheeldon, L. R. & Lahiri, A. 2002. The minimal unit of phonological encoding: prosodic or lexical. *Cognition* 85, B31-B41.

[32] Wickham, H. 2018. *ggplot2* [R package]. v.3.1.