VERBAL AND SPATIAL WORKING MEMORY LOAD HAVE SIMILARLY MINIMAL EFFECTS ON SPEECH PRODUCTION

Ogyoung Lee and Melissa A. Redford

University of Oregon
ogyoung@uoregon.edu and redford@uoregon.edu

ABSTRACT

The goal of the present study was to test the effects of working memory on speech production. Twenty American-English speaking adults produced syntactically complex sentences in tasks that taxed either verbal or spatial working memory. Sentences spoken under load were produced with more errors, fewer prosodic breaks, and at faster rates than sentence produced in the control conditions, but other acoustic correlates of rhythm and intonation did not change. Verbal and spatial working memory had very similar effects on production, suggesting that the different span tasks used to tax working memory merely shifted speakers' attention away from the act of speaking. This finding runs contra the hypothesis of incremental phonological/phonetic encoding, which predicts the manipulation of information in *verbal* working memory during speech production.

Keywords: working memory, cognitive load, speech production, prosody

1. INTRODUCTION

We know that the selection of grammatical form and lexical content (i.e., language planning) involves working memory [7, 12]; the system used for the temporary storage and manipulation of information. The type of information processed in working memory defines different components of the system [1]. Tasks that involve the manipulation of spatial relations recruit spatial working memory; those that involve the manipulation of linguistic information recruit verbal working memory. Previous research shows that verbal working memory capacity predicts individual differences in phrasing [10], and is thus implicated in speech planning. The question addressed in the present study is whether verbal working memory is similarly implicated in speaking.

In some models of speech production, planning and speaking are interleaved (e.g., [3, 8, 13]). Sentences are prosodified and the segmental content incrementally specified for execution by the motor system. Under such a model, verbal working memory is important to speaking because one must hold in memory the to-be-articulated sentence while the suprasegmental and segmental phonetics are elaborated. Because working memory is capacity limited [4], it can be overloaded and thus made ineffective. In the present study, we used this feature of working memory to test for specific effects of *verbal* working memory on speaking and compared these effects to the effects of mere load by taxing spatial working memory in a separate condition. Speech elicited under the different load conditions was compared to that elicited under control, no-load conditions. Because there is substantial phonetic evidence to suggest that speech planning occurs at the level of the prosodic phrase (see, e.g., [5, 6, 10, 13]), our focus was on how the different types of cognitive load affected suprasegmental patterns.

2. METHODS

2.1. Participants

Participants were 20 college-aged adult native speakers of American English (7 females), recruited from the Psychology and Linguistics Human Subjects Pool at the University of Oregon. Participants reported normal hearing, speaking and reading abilities and no history of speech-language therapy.

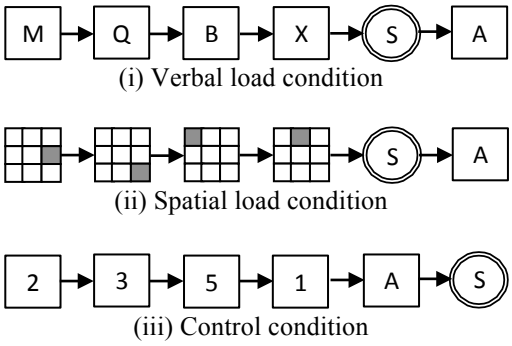
2.2. Materials

We manipulated syntactic structure in order to elicit different prosodic structures. Thirty-two sentences were designed around four complex sentences, all with dependent relative clauses. Relative clauses were either embedded in the middle of the matrix clause or appended to the end of it; for example, *the smart shy boy that liked the quiet girl cut the cake* versus *the sly gray wolf bit the sheep that wore the gold bell*. Relative clause type was either subject-extracted (as in the prior examples) or object-extracted; for example, *the fat black cat that the mad dog hurt climbed the tree* and *the swank rich man bought the paint that the young girl chose*. As in the examples given here, phrase length was controlled. Each sentence consisted of 12 monosyllabic words: 3 definite determiners; 3 adjectives; 3 nouns; 2 verbs; and 1 relativizer. There were 8 sentences for each syntactic structure.

2.3. Task

The experimental task manipulated cognitive load and the type of working memory taxed (load type). The basic framework was a complex span task (as described in [14]). A speaker was required to remember a sequence of letters (verbal load) or spatial locations (spatial load) while reading one of the stimulus sentences from a computer monitor. After reading the sentence, the speaker was then required to choose, from among a set of 8 options, the correct answer. In the control conditions, participants were presented with a sequence of numbers and asked to choose the correct sum from among 8 options before reading the stimulus sentence aloud. From a procedural perspective, this meant that the primary elicitation task came either between the serial presentation of to-be-remembered items and recall (load condition) or after recall (control condition). Figure 1 provides an illustration of the task.

Figure 1: The task manipulated cognitive load and the type of working memory taxed. Consonants, grids, or numbers were presented one at a time in the order in which they were to be remembered (or collated or summed). The sentence (S) for elicitation was presented either before or after the 8 options used to test recall (A).



During presentation, each letter / spatial location / number remained on the computer monitor for 800 milliseconds. Each sentence to be produced was displayed for 8 seconds as were the 8 response options. Importantly for the purpose of taxing verbal working memory, all the letters were consonants, in non-permissible sequences according to English phonotactics, and none of the sequences formed acronyms. In addition, the response options given during recall were highly confusable: every option repeated at least part of the correct sequence. The difficulty of the task ensured that working memory would in fact be taxed. Participants did seem to find the task challenging, and reported thinking that the primary goal of the experiment was to assess working memory and that the elicitation portion was secondary to this goal (i.e., used as a "distractor task"). The response scores were consistent with this feedback from participants.

2.4. Procedure

Prior to the main experimental task, participants were given as much time as they needed to read through the 32 sentences. We emphasized that participants have confidence in their comprehension of all sentences before they began the main task. This familiarization procedure was intended both to control for effects of language planning and to encourage a meaningful prosodification of the sentences during elicitation.

Next, participants were provided with practice in the main task, which was comprised of both the span and elicitation tasks. This practice session used a simple sentence, different from the complex sentences that were the focus of the study.

Once participants had completed the practice session, they proceeded to the main task, which was blocked by cognitive load condition and load type. Four sentences from each sentence structure were assigned to each load type so that participants produced the same 16 sentences in the load and associated no-load control condition. The set of 16 sentences produced under verbal or spatial load was counterbalanced across participants. Participants were also randomly assigned to either complete the spatial and associated no-load control task then the verbal load and associated control task or vice versa. The order in which the load and control task was completed was also counterbalanced across participants. The fixed letter / grid sequences were also randomly paired with the 16 sentences, and all was randomly presented to the participants.

Participants' speech was digitally recorded for later analysis using a Marantz PMD660 and Shure ULXS4 standard wireless receiver and lavaliere microphone. The microphone was attached to a hat that participants were given to wear. The entire experiment took no more than 90 minutes to complete.

2.5. Coding and measurement

The data from one participant was excluded at the outset because the majority of his productions were not recorded due to technical difficulties. The first author listened to the remaining 1,216 sentences (= 19 participants * 32 sentences * 2 load types), and determined that 54 additional items should be excluded from analysis due to a non-linguistic disruption during production. The remaining 1,162 sentences were then measured and coded by means of *Praat* software. 278 sentences were produced

with at least one phonological or lexical error (insertion, deletion, or replacement). Of the 884 sentences that were correctly produced, 526 sentences were load/no-load matched sentential pairs produced by the same speaker.

Phrasal breaks were perceptually identified and marked in all 884 sentences that were correctly produced. Phrase breaks were identified based on repeated listening and visual inspection of the utterance waveform for evidence of pausing and changes in intonation and temporal patterning. The spoken stretches were also marked off from silent pauses. This allowed us to calculate articulation rate in syllables per second (i.e., speech less pauses) for each of the sentences.

The 526 matched sentences of the 884 correctly produced sentences were also segmented to allow for the extraction of vocalic durations and median F0s from each rhyme in every sentence. We followed the segmentation procedures outlined by Low, Grabe, and Nolan [9] and used the temporal correlate of speech rhythm that they introduced, namely, the mean normalized sequential variability in vowel durations (nPVI). We also calculated (i) a normalized measure of F0 variation, which was the standard deviation in F0 across rhymes in a sentence divided by the mean F0 for that sentence, and (ii) the pitch range for each sentence.

2.6. Analyses

The number of sentences with at least one error was summed within a speaker and within condition (load vs. control), load type (verbal vs. spatial), relative clause type (subject-extracted vs. object-extracted), and relative clause location with respect to the matrix clause (middle vs. end). These sums then became the dependent variable in an analysis that used generalized linear mixed effects modelling (with a log link function) to assess the effects of the fixed factors on error production. Generalized linear mixed effects modelling was also used to assess the effects of the same fixed factors on the other suprasegmental measures. A linear link function was used with the normally distributed ratio data. In each case, the dependent variable represented a per speaker sum or average value that was calculated across the 4 sentences within each cell of the design.

3. RESULTS

The analyses investigated the effects of cognitive load (load condition and load type) and sentence structure (relative clause location and type) on all measures taken: (i) error rates; (ii) prosodic breaks; (iii) articulatory rate; (iv) the mean normalized sequential variability in vowel durations; (v) the normalized measure of F0 variation; (vi) pitch range.

The results indicated a significant effect of load, $F(1, 290) = 11.78$, $p < .001$, and of load type, $F(1, 290) = 5.31$, $p = .022$, on number of sentences produced with at least one error. The interaction between relative clause type and location with respect to the matrix clause was also significant, $F(1, 290) = 4.81$, $p = .029$. The direction of these effects is discernible in Table 1, which shows the cumulative number of sentences produced with error(s) by each cell in the design. As might be expected, more sentences were produced with (an) error(s) when participants were under working memory load. Also, taxing verbal working memory induced speakers to make somewhat more errors than taxing spatial working memory.

Table 1: Cumulative number of sentences with error by condition (load = L vs. control = C), load type (verbal = V vs. spatial = S), relative clause (RC) location with respect to matrix clause (middle = M vs. end = E), and RC type (subject-extracted = Sbj vs. object-extracted = Obj).

RC		L		C		
		V	S	V	S	Total
M	Sbj	25	22	20	14	81
	Obj	17	16	13	10	56
E	Sbj	19	18	13	11	61
	Obj	33	20	18	9	80
Total		94	76	64	44	278

The next set of analyses investigated differences in the measured perceptual and acoustic correlates of rhythm and intonation as a function of condition, load type, relative clause type and location. Recall that these analyses include only measures from sentences that were produced correctly (i.e., with no phonological or lexical errors).

The effect of condition was significant on the mean number of perceived prosodic breaks produced by a speaker, $F(1, 109) = 6.61$, $p = .012$. In particular, participants produced fewer prosodic breaks when speaking under working memory load than in the no load, control condition. Participants also produced more prosodic breaks in sentences with object-extracted relative clauses than in those with subject-extracted relative clauses, $F(1, 109) = 4.75$, $p = .032$.

An analysis on articulation rate was consistent with this finding: participants spoke faster in the load condition than in the control condition, $F(1, 286) = 4.40$, $p = .037$. Although there was no effect of load type on articulation rate, there were effects of structure: participants produced sentences with subject-extracted relative clauses faster than those

with object-extracted relative clauses $F(1, 286) = 7.67$, $p = .006$; and they also produced sentences with relative clauses at the end of the matrix clause faster than those with relative clauses that were in the middle of the matrix clause, $F(1, 286) = 15.882$, $p < .001$.

There was no effect of condition or of load type on the temporal correlates of spoken rhythm, although there was a difficult-to-interpret interaction between load type and relative clause location, $F(1, 267) = 6.50$, $p = .011$, on the mean temporal variability of vowel durations within a sentence.

Effects of sentence structure were also observed in the analysis of mean normalized sequential variability in vowel durations (nPVI); specifically, there were main effects of relative clause type, $F(1, 267) = 51.87$, $p < .001$, and location, $F(1, 267) = 16.78$, $p < .001$, and an interaction between these two factors, $F(1, 267) = 4.78$, $p = .030$. The pattern was for greater variability across sentences with subject-extracted relative clauses than in those with object-extracted relative clauses, and for greater variability when the relative clauses occurred at the end of the matrix clause rather than in the middle of these. The effect of relative clause location was stronger for sentences with subject-extracted relative clauses than those with object-extracted relative clauses.

As with the temporal correlates of prosody, there was no effect of condition or load type on the F0 correlates. Instead, only sentence structure mattered. There was a significant effect of relative clause type on the normalized measure of F0 variation across the sentence, $F(1, 271) = 6.13$, $p = .014$, and on pitch range across a sentence, $F(1, 271) = 4.47$, $p = .035$. The effect of relative clause location with respect to the matrix clause was also significant for both F0 variation, $F(1, 271) = 14.85$, $p < .001$, and F0 range, $F(1, 271) = 13.94$, $p < .001$. Participants produced sentences with subject-extracted relative clauses with more F0 variability and a greater F0 range than those with object-extracted relative clauses. Variability was also higher and F0 range greater in sentences with relative clauses at the end of the matrix clause than in those with relative clauses in the middle of the matrix clause.

4. GENERAL DISCUSSION

The central result from the current study is a general effect of cognitive load on speaking. Speakers made more errors when speaking under load compared to no load. They also spoke faster and with fewer prosodic breaks. The generality of this effect across load types undercuts the idea that verbal working memory is relevant to the speech production process. Instead, the results seem to be more consistent with the effects of load that have been observed in the non-language domain. For example, like the rest of us, air-traffic controllers are more prone to error when multi-tasking; however, individuals with higher working memory capacities perform more accurately under these conditions than those with lower working memory [2], and this effect is independent of general intelligence. From this we might conclude that the effects of load that we found here are attentional in nature; the present findings thus best explained as due to the cycling of attention between speaking and the span task. Cycling back and forth in this way would have reduced the overall amount of attention (and time) speakers devoted to production.

A second important result from the present study is the persistent effect of sentence structure on production. In line with differences in articulation rate, participants produced more prosodic breaks in sentences with object-extracted relative clauses than in those with subject-extracted relative clauses. They also produced sentences with relative clauses at the end of the matrix clause faster than those with relative clauses that were in the middle of the matrix clause. Effects of relative clause type and location were also observed in the analyses on the acoustic correlates of rhythm and intonation. Note that none of these effects are likely attributable to language planning, which we controlled for by familiarizing participants to the sentences prior to the main experimental task and by having participants read the sentences off of a computer monitor during elicitation.

Although the effects of structure are probably not attributable to language planning, they do represent the kinds of structure-driven differences in prosodification that we would expect from natural speech. So how might we account for this? One possibility is that speakers prosodified the sentences to themselves upon first encounter (i.e., during the familiarization period), and then used the same template when producing the sentences later during the experimental task. This possibility is consistent with a view of speech production as remembered action [11]: the speech plan is a guide for speech action, but it is not itself planned; instead, the plan represents the sequential activation of schemas that are immediately accessible to the motor system.

5. ACKNOWLEDGMENTS

The work was supported in part by a grant from the NIH, #R01HD061458. The content is solely the responsibility of the authors and does not necessarily reflect the views of the NIH.

6. REFERENCES

[1] Baddeley, A. 1992. Working memory. *Science* 255, 556-559.

[2] Colom, R., Martínez-Molina, A., Shih, P. C., Santacreu, J. 2010. Intelligence, working memory, and multitasking performance. *Intelligence* 38, 543-551.

[3] Dell, G. S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93, 283-321.

[4] Engle, R. W. 2002. Working memory capacity as executive attention. *Current Directions in Psychological Science* 11, 19-23.

[5] Ferreira, F. 1993. Creation of prosody during sentence production. *Psychological Review* 100, 233-253.

[6] Fougeron, C., Keating, P. A. 1997. Articulatory strengthening at edges of prosodic domains. *J. Acoust. Soc. Am.* 101, 3728-3740.

[7] Gathercole, S. E., Baddeley, A. D. 2014. *Working Memory and Language Processing* Psychology Press.

[8] Levelt, W. J. M. 1989. *Speaking: From Intention to Articulation* Cambridge, MA: MIT Press.

[9] Low, L. E., Grabe, E., Nolan, F. 2000. Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. *Language and Speech* 43, 377-401.

[10] Petrone, C., Fuchs, S., Krivokapić, J. 2011. Consequences of working memory differences and phrasal length on pause duration and fundamental frequency. *Proceedings of the 9th International Seminar on Speech Production (ISSP)*, Montréal, Canada.

[11] Redford, M.A. 2015. Unifying speech and language in a developmentally sensitive model of production. Resubmitted.

[12] Swets, B., Jakovina, M.E., Gerrig, R.J. 2014. Individual differences in the scope of speech planning: evidence from eye-movements. *Language and Cognition* 6, 12- 44.

[13] Shattuck-Hufnagel, S. In press. Prosodic frames in speech production. In: M. A. Redford (ed), *The Handbook of Speech Production*. Boston, MA: Wiley.

[14] Unsworth, N., Heitz, R. P., Schrock, J. C., Engle, R. W. 2005. An automated version of the operation span task. *Behavior Research Methods* 37, 489-505.