Effects of Compatible versus Competing Rhythmic Grouping on Errors and Timing Variability in Speech

Argyro Katsika  
Haskins Laboratories, USA

Stefanie Shattuck-Hufnagel  
Speech Communication Group, USA

Christine Mooshammer  
Haskins Laboratories, USA; Humboldt University of Berlin, Germany

Mark Tiede  
Haskins Laboratories, USA

Louis Goldstein  
Haskins Laboratories, USA  
University of Southern California, USA

Abstract
In typical speech words are grouped into prosodic constituents. This study investigates how such grouping interacts with segmental sequencing patterns in the production of repetitive word sequences. We experimentally manipulated grouping behavior using a rhythmic repetition task to elicit speech for perceptual and acoustic analysis to test the hypothesis that prosodic structure and patterns of segmental alternation can interact in the production planning process. Talkers produced alternating sequences of two words (top cop) and non-alternating controls (top top and cop cop), organized into six-word sequences. These sequences were further organized into prosodic groupings of three two-word pairs or two three-word triples by means of visual cues and audible metronome clicks. Results for six speakers showed more speech errors in triples, that is, when pairwise word alternation was mismatched with prosodic subgrouping in triples. This result suggests that the planning process for the segmental units of an utterance interacts with the planning process for the prosodic grouping of its words. It also highlights

Corresponding author:  
Argyro Katsika, Haskins Laboratories, 300 George Street, Suite 900, New Haven, CT 06511, USA.  
Email: argyro.katsika@haskins.yale.edu
the importance of extending commonly used experimental speech elicitation methods to include more complex prosodic patterns, in order to evoke the kinds of interaction between prosodic structure and planning that occur in the production of lexical forms in continuous communicative speech.

Keywords
speech errors, prosodic structure, rhythmic grouping, speech production planning, error elicitation methods

1. Introduction

Ever since Fromkin (1971) revived modern interest in speech errors as a source of evidence for evaluating models of language representations and speech production, investigators have searched for ways to reliably elicit errorful examples, in order to ask focused questions about the relation between phonological structure/context and error patterns. This search has been inspired by the fact that, although errors collected by listening to ongoing speech have provided a fertile source of hypotheses about the nature of the speech production process, and have imposed strong constraints on the set of possible models of this process, corpora collected in this way have a number of important limitations. As many researchers have pointed out, they are subject to potential biases, such as selective perception (Browman, 1978), selective memory, and the Procrustean effect of symbolic transcription. This last reflects the fact that the alphabetic transcription method generally used to transcribe speech error corpora does not capture certain kinds of information, such as the precise acoustic and articulatory shape of the error (and its correction, if any), or its prosody (see Goldrick & Blumstein, 2006; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; Mowrey & MacKay, 1990; Pouplier & Goldstein, 2005).

To address these problems, investigators have turned to eliciting errors experimentally in the laboratory, under conditions that support quantitative analysis of the acoustics and articulatory characteristics of both the errorful and non-errorful speech tokens. However, these investigators have faced what might be called the elicitation conundrum: errors occur relatively rarely in typical connected speech, but the tongue twisters that have been used to elicit a higher error rate experimentally are very different in their prosodic and phonological structure from the utterances that make up most of a speaker’s typical speech production. Such twisters usually consist of strings of words that contain alternating sequences of sounds, such as top cop top cop..., repeated many times and produced in a quasi-periodic temporal pattern (see, for example, Croot, Au, & Harper, 2010; Pouplier, 2003; Shattuck-Hufnagel, 1982; Wilshire, 1998). This means that the experimenter faces a choice between (a) using stimuli that elicit sufficient errors and control utterances for analysis, but invoke production mechanisms that only partially overlap with the processes of normal communicative speech, or (b) using speech stimuli that invoke the processes of normal communicative speech production, but do not efficiently elicit sufficient errors for analysis.

One approach to this problem has been to design elicitation stimuli that have some of the structural characteristics of more typical communicative speech, yet still elicit substantial (although smaller) numbers of errors. For example, Baars, Motley, and MacKay (1975) developed a method for priming errors using sequences of different two-word phrases that were phonologically but not lexically repetitive (such as *dig beets, dog bone*) to prime phonological-level errors in phrases that reversed the phonological order of the onset consonants in the priming stimuli (e.g., *barn door*); using this priming method they showed an effect of word versus non-word status on the likelihood...
that the speaker will produce the primed phonological order in an error. This method, which has also been used by later researchers (Frisch & Wright, 2002; Pouplier, 2003, 2007), involves stimuli that are somewhat closer to typical speech than the sequences of unrelated words used in many tongue twister experiments, but still have the disadvantage of invoking minimal prosodic planning.

Two further types of methodological solutions have been proposed to increase the degree to which elicitation stimuli invoke more of the processes and representations that are involved in the production of typical communicative speech. One is to embed tongue-twister-like word sequences in short syntactically well-formed phrases, comparing for example, leap note nap lute with from the leap of the note to the nap of the lute (Shattuck-Hufnagel, 1982). Results using this method showed that, just as in the word-list twisters and in speech error corpora collected from communicative speaking contexts, word onset consonants were particularly subject to interaction errors. The second approach takes the opposite tack, using typical tongue-twister-based word strings, but expanding the range of their structured alternation patterns and/or of the prosodic structures in which they are produced. By explicitly controlling and manipulating the target grouping and accent patterns resulting from the repetitively alternating target words experimenters can test the effects on error patterns of introducing prosodic aspects of speech planning in this limited way. For example, Pouplier (2003) compared different patterns of alternating accent within word pairs (e.g., top COP top COP versus TOP cop TOP cop).

Taking this one step further, Croot and colleagues (Croot, Chilko, Kember, Auton, & Barnes, 2009; Croot et al., 2010) used target word quadruples embedded in sentence contexts, such as Dash would just gaze at her, Gab could not doubt, to compare the effects of different patterns of consonant alternation within the repeated target word quadruple (e.g., onset consonant pattern /d,g,g,d/ versus /d,g,d,g/, or more generally ABBA versus ABAB). They showed that the phrasing and prominence patterns of the utterance influenced the patterns of errors in word form encoding, which suggests that speakers have planned at least some aspects of these higher-level prosodic structures at the moment when serial ordering errors among sound segments occur. Finally, in a related vein, Choe and Redford (2010) took the intonational phrasing of their speaker’s output into account in their analysis of the resulting errors, although they did not control it during elicitation.

These prosodic manipulations take advantage of the effectiveness of tongue-twister-like stimuli for eliciting substantial numbers of errors (and for providing ready comparison to non-error-inducing control utterances), while also taking important steps toward replicating at least some of the more complex processes of prosodic grouping and accenting that characterize the production of typical communicative utterances. Such methodological developments are particularly important in view of the accumulating evidence that prosodic structure governs much of systematic context-driven phonetic variation (Byrd & Saltzman, 1998; Dilley, Shattuck-Hufnagel, & Ostendorf, 1996; Fougeron & Keating, 1997; Keating, Cho, Fougeron, & Hsu, 2003). These observations have inspired models of production planning that incorporate the formation of prosodic structure from the earliest stages of production processing (Croot et al., 2009; Ferreira, 2007; Keating & Shattuck-Hufnagel, 2002; but cf. Levelt, Roelofs, & Meyer, 1999), raising important questions about how and when these evolving prosodic structures interact with other aspects of the planning process, such as segmental sequencing.

The present study extends the development of prosodically controlled elicitation methods that use tongue-twister-like sequences, by focusing on the interaction between two separate stimulus characteristics in the elicitation of sound-level errors: (a) prosodic grouping and (b) word alternation pattern. This is accomplished by directly comparing two different conditions of prosodic grouping for the same target sequence of alternating words: one in which the prosodic grouping of the words aligns with the pattern of word and segment alternation (e.g., top cop, top cop, top cop),
and another in which the prosodic grouping conflicts with the alternation pattern (e.g., top cop top, cop top cop). If the sequencing of speech segments interacts with prosodic processing during production planning, as predicted by models such as the Prosody First model of Keating and Shattuck-Hufnagel (2002) and other approaches cited above, and if a conflict between the alternation-based grouping and the prosodic grouping elicits more errors, then the error rate should increase for the conflicting condition.

Thus, the two goals of this study are to evaluate the efficacy of a new method for eliciting different prosodic groupings for a stimulus word sequence (using auditory along with visual cues), and to test for an interaction between the word sequence grouping pattern (imposed by the alternating onset consonants in the word sequence) and the prosodic grouping pattern (imposed by the cueing stimuli). The first, exploratory goal has two parts: to test whether auditory (accompanied by visual) cueing can elicit the target prosodic groupings, and to determine an optimal set of time intervals for the temporal cueing stimuli. Ideally these should elicit both the targeted prosodic structures, and a substantial number of targeted errors, while preventing speakers from ‘freezing up’ (we observed that under particularly challenging conditions, some speakers stopped speaking altogether).

If it is established that the grouping cues (i.e., auditory clicks plus visual color coding and punctuation of the visually presented target stimuli) elicit the expected prosodic grouping, we can proceed to test two hypotheses concerning the interaction between prosodic structure and word sequence structure in the speech production planning process. The first hypothesis is that planning for prosodic grouping and planning for segmental sequencing interact. As a result, when there is a conflict between the grouping suggested by the segmental sequencing (resulting from the pairwise word alternation) and the grouping suggested by the prosodic phrasing cues, the resulting speech will be less regular. We test two predictions following from this hypothesis. The first prediction is that speakers will produce more errors involving confusions between the similar alternating onset consonants /t/ and /k/ in, for example, the word pair top cop when the string of pairwise alternating words is organized into word triples than when it is organized into word pairs, due to the mismatch between word alternation and phrasing (because such errors are expected to arise from the pattern of onset alternation we refer to them as ‘predicted’ errors). The second prediction is that this organizational difference will also be reflected in greater timing variability for triples than for pairs.

The second hypothesis is that predicted errors are less likely to occur when speakers set their own production rates than when the speech production system is under metronome-driven rate pressure. We make three predictions with regard to the condition in which speakers are left to their own devices in the timing domain (e.g., when they produce stimuli in a self-paced condition as opposed to the metronome-paced condition). Firstly, they will speak more slowly, in order to produce fewer errors; secondly, there will be fewer predicted errors observed for both of the prosodic grouping conditions (triples and pairs); and thirdly, consistent with the occurrence of fewer errors in the self-paced condition, there will be less timing variability for both prosodic grouping conditions.

Results for evaluating these hypotheses will be presented in terms of duration measures based on the temporal interval between the onset stop bursts in successive words, and measures of perceptually evaluated error rates. The implications of the results from each measure for the hypotheses and predictions will be considered in Section 4.

2. Methods

2.1 Speakers

Three male and three female native speakers of American English participated in the experiment. All were naïve to the purpose of the study, were between the ages of 19 and 27, and had no
2.2 Stimuli and elicitation

The data reported here were collected as part of a larger experiment that involved a number of manipulations of the segmental sequencing of the target words, as well as a range of prosodic grouping patterns. Here we will report results for three segmental sequencing manipulations, specifically the alternating word pair top cop, and its control pairs top top and cop cop, and two prosodic grouping manipulations, namely, organization into word pairs versus word triples. Each of the three segmental sequencing manipulations was presented to speakers in stimuli consisting of groups of six words, in other words alternating (top cop top cop top cop), control-cop (cop cop cop cop cop cop), and control-top (top top top top top top). The alternating stimulus top cop was chosen to test articulatory interaction between consonants (t/k) in the onset position of the words, since it has been shown to elicit both perceptible and non-perceptible articulatory errors (e.g., Pouplier & Goldstein, 2005). The prosodic grouping manipulation consisted of presenting each of these six-word target repetition groups with two different grouping structures: one with pairs of words (e.g., top cop, top cop, top cop) and one with triples (e.g., top cop top, cop top cop). Henceforth, the groups of six words are referred to as groups, and pairs and triples as subgroups. Within a given group, the sequence of words was the same, but the prosodic subgrouping was different, so that for the alternating word pair, top cop, one of the grouping conditions (subgroup: pairs) aligned the prosodic structure with the word alternation pattern, while the other (subgroup: triples) presented the speaker with a conflict between the word alternation and the grouping pattern. The grouping manipulations not presented here were groups of four, two, or one words that did not create any contrast between aligning prosodic grouping and word alternation patterns, and thus could not serve the purpose of the current study. Also, an additional pair of words (i.e., pip kick) was used to test for control of segmental sequencing. This manipulation is not reported here for two reasons: firstly, due to time limitations caused by the length of the experiment, the adequate controls (i.e., pip pip and kick kick) were not included in the experimental design; secondly, the pip kick tokens were highly disfluent, precluding appropriate analysis of prosodic differences.

In order to elicit speech errors, metronome pacing was provided at a relatively rapid rate. Self-paced speech was used as a control, to evaluate the effect of timing pressure on error rates and timing variability in the resulting speech. In both metronome-paced and self-paced conditions, the stimuli were presented visually on a monitor screen that showed the target group of six words (one group) with subgroups delineated with punctuation (e.g., comma for the end of subgroups versus semi-colon for the end of groups). In the metronome-paced conditions, the time to be devoted to each subgroup was additionally cued by an audible click, accompanied by a colored frame on the visual display that moved from one subgroup to the next in time with the clicks. For pairs, the click-interval timing was 600 ms for each of the first two word pairs and 900 ms for the final pair in each repetition group of six, leaving extra time at the end of each group; for the triples, the timing was 900 ms for the first triple and 1200 ms for the second, again leaving extra time at the end of each group. The total time cued for each repetition group was thus 2100 ms, regardless of subgroup structure.1

Each of the three target word pairs (top cop, cop cop, and top top) was elicited in each of the two grouping patterns (pairs and triples) and in each of the two speaking conditions (metronome-paced and self-paced). All trials had a duration of 20 seconds, to encourage production on a single breath; trials with alternating words (top cop) were produced twice, while trials with control stimuli (top top...
top cop top, cop top cop;

**Figure 1.** An illustration of the screen during a metronome-paced trial for a top cop stimulus organized into subgroups of triples. The green frame (shown here as gray) indicates that the speaker should produce the first subgroup of the stimulus. Note the additional visual cue of the punctuation (comma after the first three-word subgroup, semi-colon at the end of the larger six-word repetition group).

top and cop cop) were produced only once, since their primary purpose was to provide a comparison of temporal structure and variability with the alternating stimuli, and it was not predicted that any audible errors would occur on these control stimuli. Thus, the portion of the experiment discussed here consisted of 16 trials: one alternating word pair (top cop) and two control word pairs (top top and cop cop) were produced in two prosodic grouping conditions (pairs and triples) and in two speaking conditions (metronome-paced and self-paced), yielding 12 trials. The conditions that involved the alternating word pair were produced twice, giving four additional trials.

On each metronome-paced trial, participants first viewed a visual display of two presentation cycles of the target repetition group without speaking. These two familiarization cycles showed the ‘visual metronome’ in the form of a transparent red-colored frame that indicated the time for production of each target subgroup (pair or triple) as it moved from one subgroup to another, in time with corresponding cycles of the auditory metronome timing clicks. This familiarization period introduced the timing and grouping cues for that trial. At the beginning of the third cycle, the moving color frame changed to green and the participant was expected to repeat the six-word target group, guided by the subgroup organization and timing indicated by the punctuation, the moving colored frame, and the auditory clicks (one per subgroup). Figure 1 illustrates a moment during the speaking part of a triples trial, showing the visual cue for the speaker to produce the first three words of the six-word target repetition group (i.e., the green frame, shown here as gray, which surrounds these words).

For the self-paced trials, the group/subgroup organization was cued only by the punctuation marks. In these trials, the visual display of the entire target repetition pattern of six words was surrounded by a blue frame (gray in Figure 2) for the duration of the trial; the frame did not move from one subgroup to another, since no explicit timing information was provided. Presentation of each self-paced stimulus lasted for 20 seconds, as for the metronome-paced trials. During this 20-second time period (i.e., while the visual display was shown), participants were instructed to start speaking when they felt comfortable with the pattern, and to stop repeating it when they ran out of breath. Figure 2 illustrates what the computer screen looked like during a self-paced trial.

### 2.3 Procedure

The experimental procedure consisted of two parts. In the first part, the production tasks were introduced and the participants practiced producing the test prosodic groupings and conditions
appropriately. They first practiced the metronome-paced trials, and then the self-paced ones. Speakers were instructed that semi-colons denoted stronger boundaries than commas (where ‘boundaries’ are what speakers produce between two constituents, such as phrases). For the self-paced trials this instruction was aided by (1) the fact that speakers were trained first with metronome-paced trials and (2) specific instructions to replicate the already-learned grouping patterns of the analogous metronome-paced trials. Participants were given as much time as they wished to go over the instructions and to practice producing the patterns. For the training, the word pair *tip kip* was used, to avoid biasing the experimental results with practice on the test words.

The second part consisted of the experimental trials, which were presented in a randomized order. Custom-developed software (Mark Tiede, Haskins Laboratories) was used to control the presentation of visual and audible stimuli, and to record speaker productions. To avoid contamination of the audio signal (recorded as one channel of the files stored for each trial), the metronome clicks were presented through a headset and stored as the second channel of recorded data. Audio was recorded in a sound-treated room using a Sennheiser shotgun microphone with a 22050 Hz sampling rate.

### 2.4 Labeling

For each recorded trial, the following annotations were applied using Praat (Boersma & Weenink, 2005).

1) The time of the release burst of the onset stop consonants (Stevens, 2002).

2) Speech errors. These were audible errors involving an interaction between the two target stop consonants in the word-onset position, here called predicted errors. In the *top cop* sequences used, the target consonants are /t/ and /k/. In these sequences, the production of an onset stop consonant in a location with a target /t/ counted as a speech error if it was heard as /k/ and vice versa. Errors were identified based on the perceptual judgments of two labelers; disagreements between the labelers were very rare, and were resolved by consensus listening by the authors.

3) The beginning and end times of each repetition group. The beginning was defined as the release burst of the first onset consonant of the pattern, and the end was marked at the offset of the nucleus vowel (i.e., the onset of the closure for the coda) of the last syllable in the repetition group (since the coda was frequently unreleased, so that the final release burst could not be used). Explicit labels were developed for tokens in which speakers interrupted their production in the middle of a six-word repetition group, and then either resumed the interrupted pattern or started a new repetition. These labels were used to select the uninterrupted groups for the durational analysis.

### 2.5 Analysis

For statistical analyses two variables were used: (1) the duration of the onset-burst-to-onset-burst interval, in other words, the interval from the burst of one word-onset stop to the next, and (2) the error rate derived from the predicted errors. In each trial, only the uninterrupted groups that speakers produced were analyzed for durational measures, and both interrupted and uninterrupted groups were analyzed for measures of error rate. In total, 2.68% of the produced groups were interrupted.

The durational analysis was conducted in terms of the burst-to-burst intervals (henceforth b-to-b intervals). The b-to-b interval consists of the duration of the voice onset time (VOT) plus the vowel
of the target syllable, plus the closure and release of the coda consonant and the following pause (if any), and the closure duration of the next word-onset consonant, which is terminated by the release burst for this onset consonant. Figure 3 illustrates the b-to-b intervals in a repetition with grouping in pairs.

This measure was chosen for several reasons. Firstly, the b-to-b interval includes the speech region that is subject to boundary lengthening. That is, phrase-final lengthening affects largely the final vowel and following coda consonant closure (cf. Turk & Shattuck-Hufnagel, 2007), and phrase-initial lengthening affects the initial onset consonant closure (cf. Tabain, 2003) as well as the pause; combining all of this material within a single measure that can indicate where a speaker is signaling a grouping boundary. An alternative interval of interest, the duration of a word per se, could not be measured consistently because the coda stop /p/ did not always have a visible stop release burst; it was missing in about half of the cases. For this reason it was also often impossible to measure the pause between words consistently as a correlate of boundary strength. Secondly, the alternative measures that are independent of an audible coda release, such as interval from onset burst to vowel offset or from p-center to p-center (determined from the intensity signal, see Cummins & Port, 1998), were found to be highly correlated with the b-to-b interval.

The b-to-b interval was used in four ways as follows:

Figure 3. The burst-to-burst (b-to-b) intervals in a single repetition of a six-word top cop group (Group 1), followed by the beginning of the following group (Group 2) in a trial with subgrouping in pairs. Note the slightly longer final b-to-b interval at the end of Subgroups 1 and 2 (b-to-b 2 and b-to-b 4), and the longer final interval at Subgroup 3 (b-to-b 6), corresponding to the boundary of the entire Group 1.
1) To test whether speakers signaled phrasing by final lengthening. For this purpose, the six intervals within a group were assigned three different levels of the factor boundary: b-to-b intervals spanning a group (word 6 in both patterns) were assigned the level GR (group), b-to-b intervals spanning subgroups (word 3 for the triples pattern and syllables 2 and 4 for the pairs pattern) were assigned SG (subgroup), and all within-subgroup words were assigned NB (no boundary). Repeated-measures analyses of variance (ANOVAs) were calculated with boundary (levels: GR, SG, NB), prosodic pattern (levels: pairs, triples) and speaking condition (levels: self-paced, metronome-paced, abbreviated as SPC and MET, respectively) as the independent variables, and b-to-b interval as the dependent variable. Pairwise \(t\)-tests with a Holm correction for multiple tests were carried for the three-level factor boundary.

2) To test whether the speakers produced the pairs versus triples prosodic groupings consistently as intended/trained. The b-to-b intervals of each of the six syllables in a group were used to calculate the percentage of correct patterns. We use percentages, and not counts, because speakers consistently produced fewer patterns in the self-paced than in the metronome-paced conditions. In order to compare the two speaking conditions to each other, we normalized. Specifically, for each group in a trial, we determined whether the following criteria were successfully met: (i) for overall success, the group-final b-to-b interval was longest, subgroup-final b-to-b intervals were shorter, and non-final ones shortest; (ii) for subgroup success, the b-to-b interval was longer at subgroup boundaries than within subgroups; and (iii) for group success, the b-to-b interval was longer at group boundaries than elsewhere. Then for each trial we determined the percent of groups that met these criteria. This percentage is reported as the success rate. The Wilcoxon signed-rank test was used to examine the effects of speaking condition (self-paced versus metronome-paced) and prosodic pattern (pairs versus triples) on success rates. We use the Wilcoxon test (and not a \(t\)-test), because our data do not show normal distribution due to the high number of 100% success rates.

3) To test whether there were differences in duration between the metronome-paced and the self-paced speaking conditions. The same analysis as in 1 was used.

4) To test whether difficulty (observed as error rates, see below) is related to temporal variability. The relative standard deviations of the b-to-b intervals (i.e., standard deviation times 100 and divided by the mean, also termed the coefficient of variation) for each of the six words within a group were computed for all the groups of a trial (yielding six relative standard deviations per trial, one for each word) and then averaged over the six. The relative standard deviation was preferred here to the standard deviation because the latter is correlated with the mean. Since the mean durations in our analyses varied widely, the relative standard deviation was analyzed. The larger this value was, the less consistently that speaker produced the pairs or triples patterns. Repeated-measures ANOVAs were calculated with relative standard deviation as the dependent variable and speaking condition (metronome-paced versus self-paced), sequencing pattern (repetition word pair top top versus cop cop versus top cop), and prosodic pattern (pairs versus triples) as the independent variables.

Predicted error rates were calculated only for the trials with alternating stimulus words (top cop), since predicted errors were undefined for the non-alternating control stimuli (top top and cop cop). Based on the observation that speakers did not always produce the same number of words for each trial, error rates were normalized to the number of words actually produced per trial. In other words the raw error count was divided by the number of words. To determine whether different
prosodic patterns (triples versus pairs) or speaking conditions (self-paced versus metronome-paced) elicited more or fewer errors, normalized error rates per trial were calculated.

3. Results

3.1 B-to-b interval

The efficacy of our new elicitation method was tested by determining whether speakers actually produced the desired target groupings. Acoustic correlates of prosodic grouping normally include some combination of boundary-related changes in duration, F0, amplitude, and voice quality (Byrd & Saltzman, 1998; Turk & Shattuck-Hufnagel, 2007, inter alia). In this study we tested whether evidence for grouping in the productions could be found in just one of these acoustic measures, namely, the duration of the interval between the release burst of successive onset consonants (e.g., in *top cop*, from the burst of /t/ to the burst of /k/), called here the b-to-b interval. As noted earlier, this interval includes the duration of the target word rhyme, the subsequent pause, and the closure of the onset consonant for the next target word. We predict that speakers will signal a group or subgroup boundary by lengthening it.

As shown in Figure 4, pooled results for b-to-b intervals show that, on average, (1) speakers produced longer durations at the boundary between each six-word repetition group than elsewhere, and (2) within each repetition, they produced longer durations at the boundary between each subgroup than at the beginning or middle of a subgroup, whether the subgroups were pairs or triples. Figure 4 shows that they were able to do this not only in the self-paced condition (SPC), where they could take as much time as they wanted to, but also in the metronome-paced condition (MET), where the timing constraints for both the larger repetition groups and the smaller subgroups of pairs or triples were specified. Repeated-measures ANOVAs showed that for both patterns and both speaking conditions the factor boundary was significant (SPC pairs: $F(2, 10) = 103.8, p < 0.001$, SPC triples: $F(2, 10) = 57.2, p < 0.001$; MET pairs: $F(2, 10) = 170.5, p < 0.001$, MET triples: $F(2, 10) = 148.70, p < 0.001$). Post-hoc pairwise *t*-tests with a Holm correction for multiple comparisons showed that all within-subgroup intervals (NB) were significantly shorter than intervals spanning subgroups (SG), and all subgroup intervals were significantly shorter than intervals spanning groups (GR). This result indicates that speakers were generally able to successfully produce the two different target subgrouping patterns of the six words of a repetition, namely, pairs of words versus triples, signaling the groupings by boundary-related duration lengthening.

In order to find out whether the b-to-b intervals were affected by the factor of speaking condition (self-paced versus metronome-paced), repeated-measures ANOVAs were calculated with b-to-b interval as the dependent variable, split by prosodic pattern. We found that speakers slowed down in the self-paced condition, and this effect of speaking condition was stronger for the subgroup and group boundaries than for the within-subgroup condition. There is a significant main effect both in triples, $F(1, 5) = 37.06, p < 0.001$, and pairs, $F(1, 5) = 42.83, p < 0.001$, with longer b-to-b intervals for the self-paced condition compared to the metronome-paced one. Speakers lengthened subgroup and group intervals in the self-paced condition significantly compared to the metronome-paced condition for both prosodic patterns (subgroup-pairs: $F(1, 5) = 97.2, p < 0.001$, subgroup-triples: $F(1, 5) = 23.5, p = 0.005$, group-pairs: $F(1, 5) = 17.9, p = 0.008$, group-triples: $F(1, 5) = 28.1, p = 0.003$). An interaction effect was found between speaking condition and boundary in both prosodic patterns (triples: $F(2, 10) = 17.61, p < 0.001$, pairs: $F(2, 10) = 10.53, p < 0.001$). Specifically, the effect is stronger for the subgroup and group boundaries than for the within-subgroup condition. Within-subgroup intervals were not significantly longer in the
The previous analysis showed that speakers on average signaled groupings by distinguishing significantly between the phrase-internal b-to-b intervals, the subgroup-spanning intervals, and the group-spanning intervals. However, these significant differences do not reflect how consistently speakers signaled these grouping differences by lengthening, because they are based on averages within a trial. Therefore, the general success of this elicitation method (as defined in Section 2.5 in terms of duration patterns in the b-to-b interval measure) was calculated by the success rate based on the number of b-to-b intervals that met the lengthening patterns induced by the presented stimuli (for details see Section 2). Table 1 shows that the percent of successful productions ranged from 67% to 100%, indicating that the speakers were generally able to produce the target patterns. The data are reported separately for pairs and triples, and for metronome-paced (MET) and self-paced conditions.
was (SPC) productions. The non-parametric paired Wilcoxon test was based on these subgroups and used in order to test whether speaking condition and/or prosodic pattern had an effect on success rate (given in %), as an indicator of how successfully the induced grouping was produced. As can be seen from Table 1 (last column), speakers were significantly more successful in the self-paced than in the metronome-paced condition in all boundary types (overall, subgroup, and group). This indicates that, when left to their own devices, speakers produced this relative timing significantly more reliably than when paced by a metronome. In the self-paced condition, speakers showed high success rates. Five out of six cells in Table 1 show 97% success or better. However, speakers had occasional difficulties signaling the differences between subgroup and larger repetition group boundaries for triples, as reflected by a success rate of 84%.

Despite the significant differences between self-paced and metronome-paced conditions, speakers were also quite successful in the metronome-paced condition, although success rates were not quite as high: five of six cells show 82% or better performance. Interestingly, as in the self-paced timing condition, the two prosodic patterns (pairs versus triples) did not always show the same pattern of success. That is, in the metronome-paced condition, speakers differentiated the group boundary (GR) (at the end of each repetition group of six words) from the subgroup boundary (SG) (at the end of each subgroup) less reliably for the triples (67%) than for the pairs (96%) ($p < 0.01$, see Table 1, row NB, SG<GR). This less-reliable difference between the subgroup boundary and the group boundary for subgrouping in arises from the fact that speakers tended to produce about the same b-to-b interval for the last word of both of the subgroups of three within a repetition group of six. That is, timing-wise, speakers tended to produce the six target words in the triples pattern as successive sequences of three, less reliably distinguishing the subgroup from the group boundary by this temporal measure. In contrast, for the pairs pattern, speakers more reliably increased the duration of the b-to-b interval for the group boundary, but this measure of boundary-related lengthening was somewhat less reliable for the subgroups boundaries between successive pairs of words within a repetition group. In other words, the distinction between subgroups and words within a subgroup was significantly less reliable for pairs than for triples ($p < 0.01$). The difference between patterns of pairs and triples in the reliability with which the b-to-b interval duration reflects the hierarchical grouping organization, despite the general success, may relate to preference differences for timing in different prosodic patterns (see Section 4). For our purposes here, however, the critical finding is that triples versus pairs subgroups were successfully differentiated; in other words, they were reliably distinct on this

---

**Table 1.** Success rates (in %) for correct patterns (pooled across speakers and items) for the four conditions (triples versus pairs in the metronome-paced condition [MET], and triples versus pairs in the self-paced condition [SPC]). The last column compares MET and SPC. The success rate termed Overall denotes the rate of patterns that were temporally correctly produced. NB<SG refers to the rate of patterns for which the burst-to-burst intervals for the subgroup were longer than for the within-subgroup condition. The success rate in the row NB, SG<GR compares the group boundaries to all others. Inequality signs (</>) indicate the direction of magnitude for significant comparisons. For more details see Section 2.5.

<table>
<thead>
<tr>
<th></th>
<th>MET</th>
<th>SPC</th>
<th>MET vs. SPC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Triangles</td>
<td>Pairs</td>
<td>Triangles vs. pairs</td>
</tr>
<tr>
<td>Overall</td>
<td>94</td>
<td>82</td>
<td>$&gt;p = 0.018$</td>
</tr>
<tr>
<td>NB&lt;SG</td>
<td>99</td>
<td>78</td>
<td>$&gt;p &lt; 0.001$</td>
</tr>
<tr>
<td>NB, SG&lt;GR</td>
<td>67</td>
<td>96</td>
<td>$&lt;p = 0.003$</td>
</tr>
</tbody>
</table>

NB: no boundary; SG: subgroup; GR: group.
measure. Other cues to grouping, including F0 marking, could also often be distinguished by the listener, but these cues were not coded here.

To summarize so far, our results confirm that the speakers were able to distinguish consistently between the different prosodic patterns. Therefore, the elicitation method reported for the first time here was successful. Moreover, our second hypothesis, that there are differences between the two speaking conditions, is supported: in the self-paced condition (SPC) speakers slowed down significantly, by lengthening the b-to-b intervals across both group and subgroup boundaries compared to the metronome-paced condition (MET); this also led to a higher success rate in producing the target prosodic grouping pattern in the self-paced condition.

3.2 Error rate

In order to test our first hypothesis, that a mismatch in segmental and prosodic planning will elicit more errors than for matched sequences, and our second hypothesis, that fewer errors will occur in the self-paced condition, predicted errors observed from the stimulus top cop were compared for prosodic patterns and speaking conditions. In the self-paced condition, only three such errors occurred; since speakers were free to produce their speech at any rate they chose, they generally spoke slowly enough to avoid making any errors. In contrast, the faster speaking rate with explicit group and subgroup timing demanded by the metronome-paced condition elicited substantial numbers of errors (40), so only the results for the metronome-paced condition will be presented here. This result confirms our second hypothesis that speakers make fewer errors when they are free to slow down in order to avoid errors.

As Figure 5 reveals, speakers produced more predicted errors (i.e., /t-k/ substitutions in word onset position) when a string of six words was produced with the triples pattern (i.e., as two groups of three words) than when that same string of six words was produced with the pairs pattern (i.e., as three groups of two words). Specifically, there were 36 errors in the triples condition (an error rate
of more than 5%) versus four errors in the pairs condition (less than 1%). Thus, our first hypothesis that a conflict between the pairwise word alternation pattern and the triples prosodic pattern would increase the rate of predicted interaction errors between the onset consonants was strongly supported, indicating an interaction between segmental sequencing and prosodic grouping pattern.

When we examine the predicted word-onset errors in triples versus pairs for each participant individually, we find a qualitative difference, in that five of the six speakers show a higher error rate for triples than for pairs (Figure 6). The one speaker who does not exhibit the pattern makes only 1.1% errors overall, so this can be seen as a ceiling performance. We conclude, therefore, that this is a reliable result.

3.3 Relative standard deviation of b-to-b interval

Our first hypothesis was that subgrouping in triples of alternating word pairs would produce both more errors of the predicted type and also greater variability in timing. The latter prediction was tested by a repeated-measures ANOVA with relative standard deviation as the dependent variable, pattern (levels: triples, pairs), speaking condition (levels: metronome-paced, self-paced), and segmental sequencing (levels: top top, cop cop, top cop) as independent factors, and speaker as a random factor. The mean relative SD for each condition is plotted in Figure 7. The factor of pattern was not significant, $F(1, 5) = 0.98, p = n.s.$, contrary to our predictions. A main effect of speaking condition was shown, $F(1, 5) = 8.3, p = 0.034$, indicating that metronome-paced stimuli consistently have greater temporal variability than self-paced stimuli, presumably reflecting the timing challenges of the metronome-paced condition. Our second hypothesis is thus confirmed. We find less variability in the self-paced condition, which is presumably attributed to the lack of errors in that speaking condition. However, it is unclear whether the increase in variability observed in the metronome-paced condition is related to the rate pressure induced by the metronome, or by the task itself, namely, speaking in time to a metronome. A separate experiment using different metronome rates is needed to address this issue. Here, the primary goal of the metronome condition, to induce errors, was successfully achieved.
Finally, the factor segmental sequencing also showed a main effect, $F(2, 10) = 5.3, p = 0.027$, indicating that the control sequence `cop cop` produced more variability than the control sequence `top top`, and just about as much variability as the alternating sequence `top cop`. Separate ANOVAs for pairs and triples showed that this was true for pairs, $F(2, 10) = 5.5, p = 0.025$, but not for triples, $F(2, 10) = 1.98, p > 0.05$. For the triples in the metronome-paced condition, the variability for the control sequence `cop cop` even tended to exceed that for the alternating sequence `top cop`. In contrast, the second control sequence `top top` showed less variability. Additional evidence for more variability in `cop cop` sequences than in `top top` is presented by our count of incomplete phrases (see Section 2.4), in other words, phrases that could not be completed due to hesitations or time pressure. The speakers in this study produced all phrases correctly for `top top`, six phrases were incomplete for `cop cop` and 13 for `top cop`. One characteristic that `cop cop` and `top cop` have in common, not shared by `top top`, is the coda#onset sequence `/-p#k-/` across a word boundary. Although it is not clear how this characteristic could increase the variability in boundary-related durations, it is interesting to note that Gao and colleagues (Gao et al., 2011), using Electromagnetic Articulography (EMA) measures of tongue position, reported somewhat greater consonant overlap across word boundaries for `/-p#k-/` sequences than for `/-p#t-/` sequences.

4. Discussion

This study established the effectiveness of a new method for eliciting the target prosodic structures, and then tested two hypotheses: that planning for prosodic grouping interacts with planning for the sequencing of segments, and that predicted segmental errors are particularly likely to occur when the speech must be produced under rate pressure. We tested these hypotheses using two types of measures: one durational (the b-to-b interval) and the other perceptual (the number of perceived errors).

Firstly, speakers successfully produced the prosodic grouping patterns defined by the task, namely, two groups of three words versus three groups of two, as measured by boundary-related lengthening of the interval defined by the time between the successive onset consonant bursts of each pair of successive words. The fact that this interval was longer for the final word of each target group shows that the stimulus presentation method was successful in conveying the target.

Figure 7. Means and standard errors of relative standard deviation of the burst-to-burst interval per syllable shown separately for triples and pairs, for metronome-paced versus self-paced conditions and for each of the four stimulus content types, averaged across the six words within each phrase.
grouping to the speakers, and that they were generally able to use duration lengthening to signal boundaries both for the larger groups of six words, and for the subgroups of pairs or triples. Consequently, we can compare the effects of the two different phrasing conditions on error and timing variation rates. Although speakers successfully distinguished between the two grouping patterns prosodically, they did not do so to the same degree in all conditions. In particular, the self-paced speaking condition triggered the elicited prosodic patterns more reliably than the metronome-paced condition. However, the self-paced speaking condition, unlike the metronome-paced one, was not successful in eliciting a significant number of speech errors, presumably due to absence of time pressure. This suggests that the metronome-paced speaking condition should be preferred to self-paced timing as a method for eliciting speech errors in prosodically complex repetitive speech.

Comparing the pairs prosodic pattern to the triples pattern shows that speakers more consistently signaled the large group-of-six boundaries in the pairs condition than in the triples condition, whereas they more consistently signaled the small subgroup boundaries in the triples condition than in the pairs condition. This difference may relate to a difference in the ‘phrasal cycle’ with which each type of phrasing was produced. The phrasal cycle, a term suggested by Cummins and Port (1998) for the timing of repeated phrases, refers to the point at which, in the larger timing cycle defined by the beats in an utterance, the next utterance begins. The timing measures reported in Figure 4 are consistent with our informal observations of differences in the way speakers prefer to produce six-word groups with pairs versus triples subgroups, when there are no timing constraints imposed by the experimenter. One way of describing this is that, for these participants, there is a tendency to prefer a four-beat organization for both the triples and the pairs patterns, but this four-beat organization favors a differing timing for the two different subgroup structures. For stimuli with three pairs of words as subgroups, the preferred timing is three beats (one on each pair of words) plus a one-beat pause, making a total of four beats, each with two sub-beats (with each sub-beat corresponding to a word in a word pair). This results in a tendency toward a longer duration at the end of each group of six (to leave room for the fourth beat), with a smaller duration increase at the subgroup boundary between successive pairs. In contrast, for stimuli with two subgroups of triples, the preferred timing seems to be one beat per word plus a one-beat pause after each group of three. This produces a tendency toward approximately the same duration pause after each group of three words (again to leave room for the fourth beat, but this time after each subgroup instead of after the entire group), thus demarcating the subgroup boundary after each group of three words more strongly than the larger group boundary after each group of six. Our metronome-paced condition may have imposed two separate constraints on the timing that speakers adopted, such that their timing performance reflected a combination of (a) the constraint to signal both the boundaries of large groups and subgroups, and (b) a preference for a four-beat organization that emphasized the subgroup boundaries for triples but the large group boundaries for pairs. An alternative account is that, because the subgrouping in triples required only a single subgroup boundary, speakers actually had more time to produce each word in the triples than in the pairs, possibly leading to less difficulty in signaling that single subgroup boundary. These possibilities, which raise interesting questions about how speakers determine the preferred inter-group timing for repetitive speech, will be investigated in future work.

The results also supported our hypothesis that planning for prosodic grouping interacts with planning for the sequencing of segments. When there was a mismatch between onset consonant alternation (which was always pairwise) and phrasing (which was either in pairs or in triples), namely, in the triples condition, the speakers’ production of the consonants was disrupted, as measured by the greater number of errors involving interactions between the two word-onset consonants, as heard by the labelers. This pattern was detected only in the metronome-paced condition,
since the number of errors produced in the self-paced condition was extremely small, presumably because speakers slowed their speaking rate to minimize the number of errors. This finding confirms the results reported by Rosenbaum, Weber, Hazelett, and Hindorff (1986) and interpreted as evidence for re-mapping between successive productions. It suggests that the rhythmic grouping pattern interacts with the segment planning processes, and that the presence of a grouping/alternation mismatch can cause difficulty with the planning process. Taken together with earlier results from, for example, Croot and colleagues (2010) and Choe and Redford (2010), which also suggest an interaction between prosodic grouping and phonological-level errors, these findings highlight the importance of both controlling for and systematically manipulating prosodic structure in experiments designed to elicit errors for the purpose of testing predictions from models of the production of typical speech. They also underline the necessity of developing error elicitation methods that come closer to requiring speakers to make use of the complex hierarchical and serial structures of more typical communicative utterances, and that lessen the quasi-periodicity of alternating pairs of words and sounds often found in tongue twister stimuli. Developing such methods will help to ensure that results obtained in elicitation experiments are relevant to models of more typical speech production.

For example, the repetitive and quasi-periodic beat timing of most tongue-twister-like elicitation stimuli that have been described in the literature, including those used in this experiment, are compatible with a planning process that creates a global timing framework for successive utterances, in other words, a phrasal cycle. The effectiveness of such repetitive patterns in eliciting errors can be understood as arising from two quite different sources: frequency-locking within a system of coupled oscillators (Haken, Peper, Beek, & Daffertshofer, 1996), or residual activation, as in the model of Sevald and Dell (1994). In the former perspective, multi-frequency oscillator ratios in motor tasks decrease in stability when they depart from simple 1:1 or 2:1 ratios. In the triples condition, the frequency ratio of the segment alternation to prosodic grouping is 2:3, while the ratio in the pairs condition is the much simpler 1:1. The repetitive patterns could lead subjects to perform the task by controlling the articulators with limit-cycle dynamics (i.e., oscillators) rather than as a sequence of discrete (point-attractor dynamic) tasks. Coupling among the induced oscillators would then be expected to lead to transitions in the oscillator patterns to dynamically simpler ones (Goldstein et al., 2007). In the latter perspective, the result can be explained by the hypothesis that residual activation increases the competition between activated segments. With more and more repetition of a word pair, more and more residual activation builds up and leads to errors (Sevald & Dell, 1994). Regardless of which of these accounts is correct (and both may be playing a role), it is unclear to what degree such an overt manifestation of the phrasal cycle plays a role in the production of more typical spoken utterances. Thus, it will be of considerable interest to determine what happens to the pattern of interaction errors as elicitation methods move further from the cyclical timing of repetitive rhythmical speech production. The results reported here suggest that one way of designing such methods, in other words, of ensuring the elicitation of enough errors to permit quantitative analysis while moving toward more typical speech timing patterns, is to create stimuli in which content repetition patterns and grouping structures are in conflict.

A third observation concerns the result that, in general, subgrouping in triples in the six-word repetition groups does not produce more timing variability (as measured in the b-to-b interval) than subgrouping in pairs. We predicted that increased error rates and increased timing variability might go hand in hand in this elicitation task, but that was not the case here. This aspect of the results suggests the need to measure and model separate aspects of the nature of ‘difficulty’ when considering speech production planning. That is, a comparison of error patterns and timing variability in utterances with temporally cycling versus non-cycling prosody may prove an important step toward understanding the role of prosodic structure in speech production planning. The results reported in this paper add to the evidence that this role is substantial.
Acknowledgments

We thank the speakers who participated in this study and the labelers, Rajdip Dhillon, Benjamin M Park, and Yue Li.

Funding

This work was supported by the NIH (grant NIDCD DC 008780).

Notes

1. The specific durations of the metronome intervals were selected after a series of pilot trials. Due to the lack of quantitative guidelines in the literature for the appropriate durations and weights of all the factors involved in prosodic phrasing, we kept our stimuli simple, assuming equal intervals for syllables and pauses in all our pilot trials.

2. Three other segmental landmarks were marked: (a) the end of an onset stop consonant’s VOT, which coincides with the vowel onset; (b) the offset of the second formant, indicating closure for the coda stop consonant; and (c) the release burst of the coda stop consonant (if visible). These landmarks were not used here.

3. Two other types of errors not reported here were also annotated. These were atypical productions of the target onset stop consonants, as, for example, when the degree of constriction was not stop-like but more like a fricative, and (b) all other errors, in other words, productions of consonants that were not related with the target onset stop consonants (e.g., where an onset /p/ was heard instead of the target /k/ in the second word, resulting in top pop instead of top cop). These types of errors are not reported, because they were very difficult to classify causing high inter-labeler inconsistency.

4. We note that the longer b-to-b intervals at subgroup boundaries and the even longer ones at group boundaries are consistent with a number of models for the mechanism of boundary-related lengthening, including the pi-gesture model proposed by Byrd and Saltzman (2003), the constituent-duration-planning view of Turk and Shattuck-Hufnagel (2007), and others. The data reported here do not distinguish between these various approaches to speech timing, in part because the prevalence of unreleased coda consonants made it impossible to determine from the acoustic signal the distribution of durational lengthening across the rhymes, pauses, and onsets that are adjacent to boundaries, about which different models make different predictions.

References


