Brain and Language 195 (2019) 104636

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Short communication

Chunking of phonological units in speech sequencing

Jennifer Segawa[a,b,1], Matthew Masapollo[b,1], Mona Tong[b], Dante J. Smith[c], Frank H. Guenther[b,c,d,*]

[a] Departments of Neuroscience and Biology, Stonehill College, 320 Washington Street, Easton, MA 02357, United States
[b] Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, MA 02215, United States
[c] Graduate Program for Neuroscience, Boston University, 72 East Concord St., Boston, MA 02118, United States
[d] Department of Biomedical Engineering, Boston University, 44 Cummington Mall, Room 403, Boston, MA 02215, United States

ARTICLE INFO

Keywords:
Speech production
Speech sequencing
Motor learning
Phonological working memory
Consonant clusters
GODIVA model

ABSTRACT

Efficient speech communication requires rapid, fluent production of phoneme sequences. To achieve this, our brains store frequently occurring subsequences as cohesive "chunks" that reduce phonological working memory load and improve motor performance. The current study used a motor-sequence learning paradigm in which the generalization of two performance gains (utterance duration and errors) from practicing novel phoneme sequences was used to infer the nature of these speech chunks. We found that performance improvements in duration from practicing syllables with non-native consonant clusters largely generalized to new syllables that contained those clusters. Practicing the whole syllable, however, resulted in larger performance gains in error rates compared to practicing just the consonant clusters. Collectively, these findings are consistent with theories of speech production that posit the consonant cluster as a fundamental unit of phonological working memory and speech sequencing as well as those positing the syllable as a fundamental unit of motor programming.

1. Introduction

A fundamental issue in the field of speech production is how speakers learn and rapidly execute sequences of phonological units (i.e., phonemes, syllables, words, phrases) as vocal tract articulations. Anyone who has ever attempted to speak a foreign language can readily attest that generating unfamiliar speech sound sequences in a fluent, coordinated, and natural-sounding way is a far from trivial skill. It is widely believed that the production of relatively long or complex motor sequences, such as the sequence of phonemes making up a sentence, involves the use of well-learned subsequences, often referred to as "chunks" (Cholin, Levelt, & Schiller, 2006; Guenther, 2016; Levelt & Wheeldon, 1994; Levelt, Roelofs, & Meyer, 1999). From a working memory perspective, chunking of frequently occurring subsequences allows such a subsequence to be treated as a single "item" in working memory, thereby reducing the processing load required to store a long sequence. From a motor control perspective, a frequently occurring subsequence could be stored as a "motor program" for producing that subsequence with rapid, highly coordinated movements that have been learned through practice (i.e., repeated production attempts).

Although researchers generally agree that subsequence chunking is a strategy utilized by the brain when producing speech sequences, no consensus has yet been reached regarding the precise nature of these chunks (see Guenther, 2016, chapter 8 for discussion). The observation of phonological segment error patterns in spontaneous speech, such as the swapping of phonemes between two consecutive words (e.g., "toff shelp" for "top shelf"), has led to the proposal that frequently produced consonant clusters are, at some level of the production planning process, treated as single chunks (e.g., Hindson & Byrne, 1997; MacKay, 1970; Shattuck-Hufnagel, 1983; Treiman, 1984; Loevenbruck, Collins, Beckman, Krishnamurth, & Ahalt, 1999); this is based largely on the observation that swapping errors often involve entire clusters moving between words (e.g., "dretter swying" for "sweater drying"). Consonant clusters have also been reported to exhibit more invariant intergestural timing than the same consonants with an intervening vowel (Loevenbruck et al., 1999). Based on considerations such as coarticulation patterns and syllable frequency effects, others have theorized that syllables are the most common chunk size for motor programs (e.g., Cholin et al., 2006; Guenther, Ghosh, & Tourville, 2006; Guenther, 2016; Kozhevnikov & Chistovich, 1965; Levelt & Wheeldon, 1994). In this view, a highly optimized sequence of movements is learned for each frequently produced syllable in the native language.

* Corresponding author at: Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, MA 02215, United States.
E-mail addresses: jsegawa@stonehill.edu (J. Segawa), mmasapol@bu.edu (M. Masapollo), djsmith@bu.edu (D.J. Smith), guenther@bu.edu (F.H. Guenther).
[1] Joint first authors.

https://doi.org/10.1016/j.bandl.2019.05.001
Received 25 April 2018; Received in revised form 10 May 2019; Accepted 15 May 2019

In a prior study from our laboratory (Segawa, Tourville, Beal, & Guenther, 2015), subjects completed a motor sequence learning paradigm in which they were trained to produce novel, meaningless speech sound sequences (monosyllabic CCVCC(C) pseudowords) consisting of consonant clusters that were either phonotactically legal in their native language of English (e.g., "blerk") or illegal in English but legal in other natural human languages (e.g., "gvasf"). Practice producing the novel utterances led to measurable performance gains (i.e., increased accuracy and reduced utterance durations) for the non-native sequences, but not the native sequences, which were relatively easy to produce even on the initial attempts. Contrasting fMRI BOLD activity patterns during production of the novel non-native stimuli with the practiced non-native stimuli yielded activity in left ventral premotor cortex (vPMC) as well as activity in a working memory network that includes posterior inferior frontal sulcus (pIFS), pre-supplementary motor area, anterior insula, and intraparietal sulcus. A meta-analysis of working memory neuroimaging studies identified left pIFS as the only portion of this network that is specialized for phonological material (Rottschy et al., 2012); in accord with this finding, the GODIVA model of speech sequencing (Bohland, Bullock, & Guenther, 2010; Guenther, 2016) posits that left pIFS is the location of a phonological working memory repository that temporarily buffers concatenated phonological units (chunks) for a planned utterance. The model further posits that this working memory circuit then sequentially activates speech motor programs, hypothesized to reside in left vPMC, in order to produce the phonological units represented in pIFS.

In the current study, we investigated the nature of the chunking process that leads to reduced processing load in phonological working memory by examining how performance gains from practicing phonotactically illegal phoneme sequences generalize to novel sequences that overlap to varying degrees with the practiced sequences. The experiment consisted of six practice blocks and two test blocks performed over two consecutive days. Four practice blocks were performed on day one, and two additional practice blocks were performed at the beginning of day two. In these sessions, speakers repeatedly produced two sets of novel CCVCC syllables (i.e., syllables that do not occur in any English words): (1) syllables that involved native (phonotactically legal) consonant clusters (*native CC*) and syllables based on non-native consonant clusters (*non-native CC*) that are phonotactically illegal in English. Based on the aforementioned results of Segawa et al. (2015), we expected to see significantly larger performance gains due to learning for the *non-native CC* stimuli than for the *native CC* stimuli (for which performance is already expected to be near ceiling at the beginning of training). The practice blocks were followed by two test blocks on day two that tested performance on four types of CCVCC syllables involving non-native clusters: (1) syllables that were included in their entirety in the practice session (*practiced CCVCC*), (2) novel syllables constructed of consonant clusters that were encountered during the practice session (*practiced CC*), (3) novel syllables containing practiced CVC "cores" but novel non-native clusters (*practiced CVC*), and (4) novel syllables containing novel non-native clusters and novel CVC cores (*novel CCVCC*). If the primary unit of motor sequence learning is the consonant cluster, we expect learning to generalize to novel syllables that contain the practiced consonant clusters but not to novel syllables involving novel clusters. In other words, we expect performance on *practiced CC* syllables to be approximately equivalent to *practiced CCVCC* syllables and better than both *novel CCVCC* syllables and *practiced CVC* syllables. Alternatively, if the primary unit of motor sequence learning is the whole syllable, we expect little or no generalization of cluster or core learning to new syllables containing these elements. That is to say, performance on the *practiced CCVCC* syllables should be better than both the *practiced CVC* and *practiced CC* syllables, with the latter two syllable types showing similar performance to *novel CCVCC* syllables (i.e., no generalization of cluster or core learning to new syllables containing these elements).

2. Results

To evaluate speech motor sequence learning, we examined changes in two measures that showed significant evidence of learning for similar speech sequences in Segawa et al. (2015): utterance durations and phoneme sequencing error rates (see Section 4 for details). Each measure was pooled and averaged within each condition and within each subject.

Evidence of learning for practiced syllables. Our first set of analyses was aimed at verifying performance improvements over the practice blocks on day one. Separate analyses of variance (ANOVAs) were performed on mean utterance durations and sequencing error rates with independent factors of sequence type (*native CC* vs. *non-native CC*) and time (first five trials vs. last five trials).[2]

The ANOVA performed on mean utterance durations revealed a significant main effect of sequence type [$F(1,10) = 67.306$, $p < 0.001$, $\eta^2 = 0.871$], such that speakers were faster at producing the *native CC* syllables ($M = 0.44$, $SD = 0.07$) compared to the *non-native CC* syllables ($M = 0.55$, $SD = 0.08$). The main effect of time [$F(1,10) = 1.010$, $p = 0.339$, $\eta^2 = 0.092$] did not reach statistical significance; however, the interaction [$F(1,10) = 10.325$, $p = 0.009$, $\eta^2 = 0.508$] was significant. To tease apart the interaction, difference scores were computed by subtracting the mean durations averaged across the first five and last five error-free trials for each sequence type. A post-hoc LSD *t*-test performed on these difference scores indicated that they were significantly larger for the *non-native CC* syllables [$M = 0.04$, $SD = 0.04$] compared to the *native CC* syllables [$M = -0.01$, $SD = 0.05$; $t(10) = -3.210$, $p = 0.009$, $d = -1.10$].

For sequencing error rates, there were significant main effects of sequence type [$F(1,10) = 45.414$, $p < 0.001$, $\eta^2 = 0.820$] and time [$F(1,10) = 37.906$, $p < 0.001$, $\eta^2 = 0.791$], as well as a significant interaction [$F(1,10) = 28.627$, $p < 0.001$, $\eta^2 = 0.571$]. To tease apart the interaction, difference scores were computed by subtracting the mean sequencing error rates averaged across the first five and last five trials for each sequence type. A post-hoc LSD *t*-test performed on these difference scores indicated that they were significantly larger for the *non-native CC* syllables [$M = 29.1$, $SD = 16.1$] compared to the *native CC* syllables [$M = 2.2$, $SD = 4.6$; $t(10) = -5.350$, $p < 0.001$, $d = -2.26$].

Overall, these results demonstrate clear evidence of learning (performance improvement) for both *native CC* and *non-native CC* syllables, with the more difficult *non-native CC* syllables showing significantly larger performance improvements.

Generalization of learning to novel syllables. Our second set of analyses were designed to examine the specificity of the motor sequence learning that occurred for syllables containing non-native consonant clusters during the practice blocks. Table 1 provides the overall frequency of each error subtype in each syllable type of the test phase. By far the most common sequencing error was the omission of one or more phonemes in the target syllable. We then conducted separate ANOVAs (with syllable sequence type as the independent factor) on the mean error rate scores for the first five trials of the test session (to minimize practice effects during the test phase) and on the mean utterance durations for the first five properly sequenced trials of the test blocks.

The left panel of Fig. 1 shows the mean utterance duration as a function of syllable type. The ANOVA showed a highly significant effect of sequence type [$F(3,30) = 4.894$, $p = 0.007$, $\eta^2 = 0.329$]. Post-hoc *t*-tests revealed that the mean durations for the *practiced CCVCC* syllables [$M = 0.54$, $SD = 0.10$] were significantly shorter than the *practiced CVC* [M = 0.59, SD = 0.14; $t(10) = -3.459$, $p = 0.006$] and *novel*

[2] An ANOVA that included the additional factor of syllable position (onset clusters vs. coda clusters) indicated no main or interaction effects involving syllable position. See *Supplementary Materials* for further details regarding analyses of syllable position effects.

Table 1
Mean error rates in the test blocks by error subtype and syllable type.

	Syllable Type			
Sequencing Errors	Practiced CCVCC	Practiced CC	Practiced CVC	Novel CCVCC
Phoneme omission	34.3	26.6	28.9	31.5
Phoneme substitution	2.3	8.8	7.4	7.7
Serial ordering error	1.1	3.7	1.1	4.0
Gross disfluency	0.3	2.0	1.1	1.1
Unrecognizable	0.0	0.9	1.4	0.3
Phoneme insertion	0.0	0.3	0.3	0.6
Non-Sequencing Errors	Practiced CCVCC	Practiced CC	Practiced CVC	Novel CCVCC
Vocoid epenthesis	20.2	21.0	29.8	23.3
Voicing assimilation	2.6	1.1	2.8	1.4

less improvement than practicing the whole syllable. The *practiced CVC* and *novel CCVCC* error rates were not significantly different from each other [$t(10) = -1.376$, $p = 0.199$].

The results of a second error rate ANOVA that included combined sequencing and non-sequencing errors[4] from Table 1 are shown in the right panel of Fig. 1. The analysis identified a significant effect of syllable type [$F(3,30) = 3.857$, $p = 0.019$, $\eta^2 = 0.278$]. Post-hoc LSD paired *t*-tests showed that the mean error rates for the *practiced CCVCC* syllables [M = 46.3, SD = 25.1] were significantly lower than the *practiced CC* [M = 59.5, SD = 22.2; $t(10) = -2.498$, $p = 0.032$], *practiced CVC* [M = 65.4, SD = 26.4; $t(10) = -2.003$, $p = 0.043$] and *novel CCVCC* [M = 66.8, SD = 23.9; $t(10) = -3.190$, $p = 0.010$] syllables. The mean error rates for the *practiced CC* syllables were not significantly lower than either the *practiced CVC* [$t(10) = -0.643$, $p = 0.535$] or *novel CCVCC* [$t(10) = -1.638$, $p = 0.132$] syllables. The *practiced CVC* and *novel CCVCC* error rates were not significantly different from each other [$t(10) = -0.420$, $p = 0.683$]. Overall this error pattern is similar to the pattern seen when only sequencing errors were considered (center panel of Fig. 1).

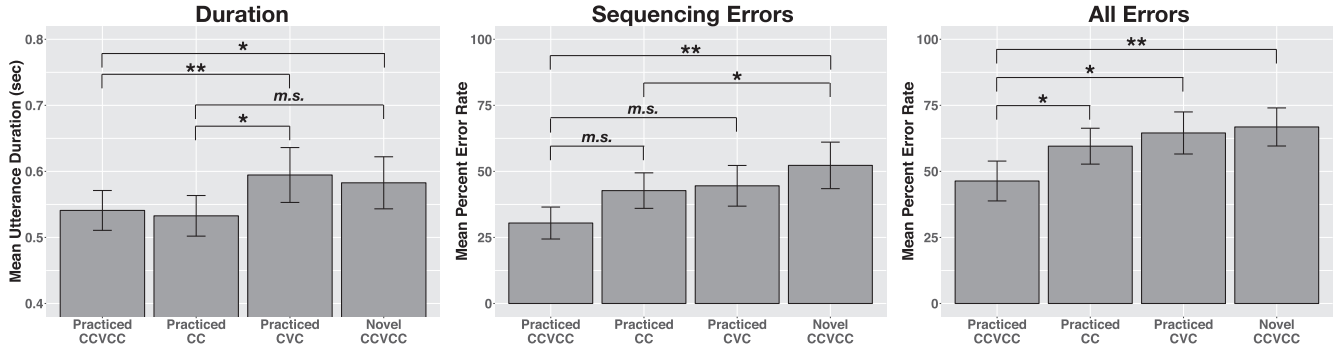

Fig. 1. Performance measures from the test blocks for *practiced CCVCC*, *practiced CC*, *practiced CVC*, and *novel CCVCC* syllables. Left panel: Mean durations of the first five properly sequenced utterances of each syllable type. Center panel: Mean percentage of sequencing errors for the first five utterances of each syllable type. Right panel: Mean percentage of total errors (sequencing and non-sequencing) for the first five utterances of each syllable type. Abbreviations: ms = marginally significant ($p < 0.1$); * = significant ($p < 0.05$); ** = significant ($p < 0.01$).

CCVCC [$M = 0.58$, $SD = 0.13$; $t(10) = -2.822$, $p = 0.018$] syllables, but not the *practiced CC* [$M = 0.53$, $SD = 0.10$; $t(10) = 0.564$, $p = 0.585$] syllables. Durations for the *practiced CC* syllables were also significantly faster than the *practiced CVC* [$t(10) = -2.548$, $p = 0.029$] and marginally faster than the *novel CCVCC* [$t(10) = -1.863$, $p = 0.092$] syllables, demonstrating a performance improvement from practicing the CC portions of the CCVCC syllables. The *practiced CVC* and *novel CCVCC* durations were not significantly different from each other [$t(10) = 0.683$, $p = 0.510$]. In sum, improvements in duration for practicing a consonant cluster entirely generalized to novel syllables with that cluster.

An ANOVA on sequencing error rate[3] (center panel of Fig. 1) showed an effect of sequence type [$F(3,30) = 4.628$, $p = 0.009$, $\eta^2 = 0.316$]. Post-hoc *t*-tests revealed that the mean sequencing error rates for the *practiced CCVCC* syllables [$M = 30.4$, $SD = 42.7$] were significantly lower than the *novel CCVCC* [$M = 52.2$, $SD = 29.2$; $t(10) = -3.425$, $p = 0.006$] syllables and marginally lower than *practiced CC* [M = 42.7, $SD = 22.2$; $t(10) = -2.218$, $p = 0.051$] and *practiced CVC* [$M = 44.5$, $SD = 25.4$; $t(10) = -1.985$, $p = 0.075$] syllables. While sequencing error rates for the *practiced CC* syllables were not significantly lower than the *practiced CVC* [$t(10) = -0.281$, $p = 0.785$], they were significantly lower than the *novel CCVCC* [$t(10) = -2.345$, $p = 0.041$] syllables, suggesting some minor improvement from practicing the CC portion of the CCVCC syllables but

3. Discussion

Current models of language and speech production commonly propose that speakers produce complex or extended sequences of speech movements by parsing them into shorter well-learned strings of movements, or "chunks". Yet there is no consensus on the precise nature and size of the chunks that play a role in the programming of speech movements. To begin to identify these chunks, the current research investigated generalization of movement chunking from training to transfer utterances. In keeping with our prior study (Segawa et al., 2015), we found greater performance gains with practice for novel syllables containing non-native consonant clusters compared to novel syllables containing consonant clusters from the native language. Furthermore, we found that speed improvements achieved by practicing a non-native consonant cluster in one syllabic context generalize fully to novel syllables that contain the practiced cluster. Specifically, we found speed gains for syllables that contained practiced consonant clusters compared to syllables in which only the CVC portion was practiced, whereas no speed difference was found between novel syllables containing practiced consonant clusters compared to syllables that were practiced in their entirety. Together, these findings support the concept of the consonant cluster as an important unit of chunking at some level

[3] An ANOVA that included the additional factor of syllable position found a marginally significant main effect of syllable position [$F(1,30) = 4.787$, $p = 0.054$, $\eta 2 = 0.324$] in which more errors occurred on coda clusters ($M = 30.3$; $SD = 25.2$) than onset clusters [$M = 21.7$; $SD = 17.3$]. See *Supplementary Materials* for further details regarding analyses of syllable position effects.

[4] An ANOVA that included the additional factor of syllable position indicated no significant effect of syllable position. See *Supplementary Materials* for further details regarding analyses of syllable position effects.

of the speech production hierarchy (e.g., Hindson & Byrne, 1997; Loevenbruck et al., 1999; MacKay, 1970; Shattuck-Hufnagel, 1983; Treiman, 1984). Also consistent with this account are recent findings that improvements in cluster production generalize from words used in training to untrained words that also contain those clusters in typical speakers and some speakers with apraxia of speech (AOS; Buchwald, Gagnon, & Miozzo, 2017; Buchwald et al., 2019).

Based on our duration results, it is tempting to conclude that the brain learns optimized speech motor programs for consonant clusters but not for full syllables. However, our analysis of sequencing error rates found significantly lower error rates when the full syllable was practiced compared to when only the consonant clusters were practiced. Thus, timing-related aspects of speech sequencing, for which the consonant cluster is a crucial unit (or possibly the phoneme-to-phoneme transition, since novel clusters involve novel consonant-to-consonant transitions), appear to be somewhat dissociated from the processes responsible for sequentially activating the proper motor gestures for the sequence, for which the syllable is the crucial unit.

Although it is possible that the brain acquires a stable motor programming unit for the cluster as well as the full syllable, it is difficult to account for why only some learning gains, namely duration reductions, transfer to the same cluster in a different phonetic context, whereas both error reductions and duration reductions are seen when the full syllable is practiced. Fig. 2 provides a possible account of these findings, based on the GODIVA model of speech sequencing (Bohland et al., 2010; Guenther, 2016). This account is consistent with the view of the consonant cluster as an important phonological unit (Hindson & Byrne, 1997; MacKay, 1970; Shattuck-Hufnagel, 1983; Treiman, 1984) and the syllable as an important unit of motor programming (Guenther et al., 2006; Guenther, 2016; Kozhevnikov & Chistovich, 1965; Levelt & Wheeldon, 1994). Each panel schematizes a simplified sequencing network consisting of a phonological working memory and a motor program repository for a syllable type in our experiment. Panel A schematizes production of a novel, non-native CCVCC syllable ("gvusb") the first time it is encountered. Since the entire syllable, including consonant clusters, is new to the speaker, each phoneme must be represented individually in the phonological working memory in left pIFS, and projections from working memory to the motor system must sequentially activate motor programs for the individual phonemes (indicated by capital letters) located in left vPMC and/or bilateral ventral primary motor cortex (vMC). Panel B represents the situation when producing "gvusb" after practicing syllables containing "gv" and "sb" but not the entire syllable "gvusb", as in our *practiced CC* condition. Practice is hypothesized to lead to two changes in the network: (1) the cluster "gv" is now represented by a single node in phonological working memory, and (2) motor programs now exist for "gv" and "sb" in left vPMC in addition to the individual phoneme motor programs in bilateral vMC. The network now has to concatenate and sequentially activate only three motor programs, one for each cluster and one for the vowel. The reduced sequencing load (including reduced working memory load) results in the faster performance and decreased error rates seen for the *practiced CC* condition compared to the *novel CCVCC* condition in our experiment. Panel C illustrates the situation when the entire syllable "gvasb" has been practiced. In this case, in addition to motor programs for the individual phonemes and consonant clusters, a motor program exists for the entire syllable in left vPMC, further reducing the sequencing load and decreasing the sequencing error rate.

According to this account, two different types of learning are occurring in our study: (1) phonological representations for the new consonant clusters are learned, and (2) optimized muscle activation patterns are learned for consonant clusters and full syllables. Based on our generalization results, we hypothesize that the primary performance benefit of the first type of learning in our paradigm is increased movement speed, whereas the latter learning process leads to fewer articulation errors. The account summarized in Fig. 2 also suggests that the chunks utilized in working memory do not necessarily correspond in a one-to-one fashion with those used for motor programs, thereby reconciling the notion of consonant cluster chunks derived from our duration results and studies of phonological segment errors (which occur when multiple words are simultaneously active in working memory; e.g., Shattuck-Hufnagel, 1983) with the notion of the syllable as a fundamental chunk size for motor programs. Because this account differentiates phonological and motoric learning processes, it may help explain why some individuals with AOS, namely those with primarily motoric deficits, improve their productions of consonant clusters with practice, while individuals with AOS who display primarily phonological deficits do not (Buchwald et al., 2017). Further research is necessary to verify or refute these predictions.

Finally, it is important to note that there are several limitations to the current study. First, because a given consonant cluster was used in either onset or coda position but not both, we are unable to investigate possible differences in how onset clusters and coda clusters are processed (cf. Byrd, 1996; Fowler, Treiman, & Gross, 1993; Kirk & Demuth, 2005) or whether motor learning generalizes to different positions within the syllable (i.e., whether consonant clusters learned as onsets generalize to the coda position or vice-versa). Moreover, only two performance measures – duration and error rate – were used to assess whether learning had occurred. These measures provide a limited window on motor learning: they cannot capture movement characteristics frequently associated with well-learned motor programs, such as increased coarticulation and reduced effort.

4. Material and methods

Subjects: Eleven subjects (6 males, aged 18–24 years, mean age = 19.7 years [*SD* = 2.2]) were recruited from the greater Boston area and paid for participating in two testing sessions. An additional four subjects were run by excluded from analysis because they did not undergo learning of the non-native CCVCC syllables (see Section 2). Subjects reported normal (or corrected-to-normal) vision and no history of hearing, speech, language or neurological deficits. All were native speakers of American English with no previous experience with any of the languages used in the stimulus creation (see following text).

Stimuli: The speech stimuli consisted of ten sets of CCVCC syllables

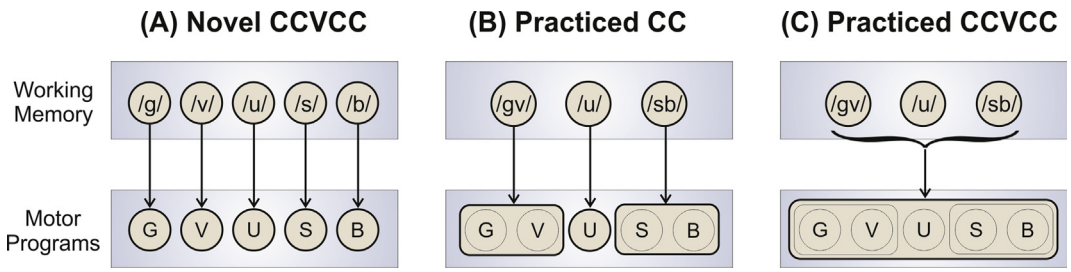

Fig. 2. Possible account of the current experimental results. Each panel represents a simplified sequencing network, consisting of a phonological working memory stage and a motor program stage, for one stimulus type in the current study. See text for details. Abbreviations: L = left; pIFS = posterior inferior frontal sulcus; vMC = ventral motor cortex; vPMC = ventral premotor cortex.

Table 2
International phonetic alphabet (IPA) transcription and orthography for experimental stimuli used to elicit the *native* (left) and *non-native* (right) target onset and coda clusters (underlined). *Non-native* sets to the left and right of each other constitute each other's *practiced* CVC sequences. Sets within the same quadrant constitute each other's *practiced* CC sequences. Sets from a diagonal quadrant constitute the *novel* CCVCC sequences (see text for explanation).

Target CCVCC Syllables								
Native			Non-Native					
Set			Set			Set		
	IPA	Orthography		IPA	Orthography		IPA	Orthography
1	blɚɹk	BLERK	3	zdɛtʃb	ZDECHB	7	bdɛtʃk	BDECHK
	flɪsk	FLISK		ʃkIzg	SHKIZG		zkIzf	ZKIZF
	gɹalv	GRALVE		fʃapf	FSHAPF		kʃæpk	KSHAPK
	pɹʌnj	PRUNGE		gvʌsb	GVUSB		zvʌstʃ	ZVUSCH
2	dɹalf	DRALF	4	fʃIzg	FSHIZG	8	kʃIzf	KSHIZF
	fɹʌmp	FREMP		gvʌtʃb	GVUCHB		zvʌtʃk	ZVUCHK
	plɚɹθ	PLIRTH		ʃkɛpf	SHKEPF		zkɛpk	ZKEPK
	tɹʌlp	TRULP		zdæsb	ZDASB		bdæstʃ	BDASCH
			5	dzukf	DZUKF	9	vzʌkp	VZUKP
				tfɛʃtʃ	TFESHCH		gfɛʃp	GFESHP
				ʃkætk	SHGATK		tgætp	TGATP
				kpImtʃ	KPIMCH		fpImʃ	FPIMSH
			6	ʃgɛkf	SHGEKF	10	tgɛkp	TGEKP
				kpæʃtʃ	KPASHCH		fpæʃp	FPASHP
				dzʌtk	DZUTK		vzʌtp	VZUTP
				tfImtʃ	TFIMCH		gfImʃ	GFIMSH

(four syllables/set). As shown in Table 2, two of the sets contained *native* onset (word-initial) and coda (word-final) clusters, and the other eight sets contained *non-native* onset and coda clusters. In the *native* sets, the onset and coda clusters occur readily in English; in the *non-native* sets, the clusters do not readily occur in English, but do occur in some other language (see Segawa et al., 2015, for further details regarding stimulus creation). A given cluster was used either in onset position or coda position, but not both. None of the subjects had prior experience with any of the languages in which these consonant clusters are legal.

To create the prompts for the elicited production task, a female native speaker of American English was recorded producing the ten sets of syllables. The model speaker was phonetically trained and had previously practiced producing the sequences until each stimulus could be executed fluently (i.e., without vowel epenthesis or phoneme omissions, swaps, or substitutions). Since we were concerned with learning of non-native phonotactics rather than sub-phonemic allophonic details, productions were not judged on how natural they sounded in the languages from which they were derived. All recordings took place in a sound-attenuated booth. Speech was recorded directly to a computer using Audacity® software (Version 2.0.3, Audacity Team) via a microphone (Samson C01U studio condenser) connected to a pre-amplifier (44.1-kHz sampling rate, 32-bit quantization). The speaker recorded multiple randomized repetitions of each token. From these repetitions, one instance of each token was selected on the basis of acoustic similarity in voice pitch (f_0) to the other stimuli in the set. Using Praat software (Boersma & Weenink, 2019), all recorded tokens were matched for peak intensity and duration (i.e., 480 ms) without changing f_0. Finally, because several of the non-native clusters disagree in their voicing specification (e.g., "gvusb"; see Table 2), a phonetically trained coder verified on the basis of auditory evaluation that the mixed voicing distinctions were indeed produced by the model speaker and present in the stimuli. Similar procedures were used to code whether subjects also produced the target mixed voicing clusters accurately.

Experimental design: Subjects completed six practice blocks and two test blocks over two consecutive days. Practice and test blocks were identical except for the stimuli used. In the practice blocks, subjects repeatedly produced the syllables from one of the *native* sets (Table 2, left) and one of the *non-native* sets (Table 2, right) in pseudorandom order. Each syllable was produced 10 times in each session, for a total of 80 productions per block. After completing six practice blocks (four on day one and two on day two), subjects completed two test blocks involving only non-native stimuli that fell into the four aforementioned categories: (1) *practiced CCVCC* syllables that were encountered in the practice sessions, (2) *practiced CC* syllables that contained non-native clusters that appeared in the practice sessions but in novel syllables, (3) *practiced CVC* syllables that included CVC "cores" that were encountered during practice but novel non-native consonant clusters, and (4) *novel CCVCC* syllables consisting of consonant clusters and CVC cores that were not encountered in the practice session. Here, each syllable was produced four times for a total of 64 productions in each test block.

Subjects were randomly divided into one of eight experimental groups, and the 32 *non-native* syllables (divided up into eight sets; Table 2, right) were used with equal frequency across subjects and groups. The two sets of *native* syllables were also counterbalanced across subjects. Subjects were seated in a chair in front of a laptop (Lenovo ThinkPad X61s) computer screen in a sound-treated laboratory room that was dimly lit. The auditory stimuli were presented over headphones (Behringer, HPM1000) at a comfortable listening level and recorded with a Samson (Hauppauge, NY) C01U USB studio condenser microphone connected to the computer via a MOTU microbook audio interface. Utterances were recorded using MATLAB (MathWorks Inc., Natick, MA).

Each trial consisted of the following sequence of events. First, the orthographic display of a given syllable (as shown in Table 2) appeared in the center of the screen along with its corresponding auditory prompt. Subjects only heard each prompt once. Then, depending on the trial, 500–1000 ms after the offset of the auditory presentation, a tone was presented for 50 ms (i.e., the onset of the tonal stimulus was randomly jittered between 500 and 1000 ms). This tone served as a go signal for the subject to repeat the token. Utterances were recorded for 1500 ms. Syllables were randomized across trials. The combination of the orthographic and auditory presentations was necessary because previous work has shown that listeners tend to perceive non-native consonant clusters as epenthesized disyllabic sequences (e.g., Berent, Steriade, Lennertz, & Vaknin, 2007; Dupoux, Kakehi, Pallier, Hirose, & Mehler, 1999; Dupoux, Parlato, Frota, Hirose, & Peperkamp, 2011).

Moreover, other research (Davidson, 2010) that directly examined the effects of stimulus input modality (audio only vs. audio and text) on speakers' ability to produce non-native consonant clusters found that the presence of text led to an improvement in overall task performance.

Subjects were instructed to repeat the stimulus as heard in the auditory prompt as quickly and accurately as possible, while making sure to produce all of the segments seen in the orthographic display. They were also instructed to attempt to eliminate any vowel-like insertions between consonants within a cluster, a common response when producing novel illegal consonant clusters (cf. Davidson, 2006). Several familiarization trials with experimenter feedback were included at the start of the experiment to confirm that subjects understood the task instructions and were able to perform the task. The *native* and *non-native* sequences used during these initial practice trials were not used at any point in the rest of the study.

Data processing. We used custom MATLAB software to perceptually rate and acoustically measure onsets and offsets of syllables by viewing the waveform and spectrogram and listening to the audio files. Each utterance was coded by a trained phonetician, blinded to experimental condition, for eight possible error subtypes: (1) gross disfluency (i.e., trials in which a subject omitted, repeated, or restarted an utterance); (2) unrecognizable from target; (3) phoneme deletion/omission; (4) phoneme insertion (i.e., one or more segments were added); (5) phoneme substitution (other than voicing assimilation of consonants in a cluster); (6) incorrect ordering of phonemes; (7) vocoid epenthesis (i.e., truncated vowel-like sounds between consonants within a cluster, evidenced by periodic peaks and a visible second formant in the spectrogram; cf. Wilson, Davidson & Martin, 2014), and (8) voicing assimilation between consonants within a cluster. Voicing assimilation was treated as a separate category from other phoneme substitution errors because the native English speakers in the current study were unfamiliar with clusters involving consonants that disagree in voicing (a very rare property cross-linguistically, but one possessed by several of the non-native CC stimuli). The lack of such clusters in English may, over the course of development, lead to a decrease in subjects' abilities to hear the voicing distinction in the target stimulus (e.g., Hallé, Segui, Frauenfelder, & Meunier, 1998; Pitt, 1998) and/or produce voicing distinctions within a cluster. Together with vocoid epenthesis errors, we consider these errors to be errors in fluency (in the sense of sounding somewhat different than a native producing the clusters, as if with a strong foreign accent) rather than errors in phoneme sequencing. Since we are primarily interested in the neural mechanisms underlying speech sequencing, our primary focus herein was on sequencing errors (subtypes 1–6). Mean error rates for each subject were calculated as the percentage of trials that contained one or more errors. For each production containing no sequencing errors, utterance onset and offset were automatically labeled based on sound pressure level thresholds, then hand-checked.

5. Statement of significance

The precise nature of the phonological working memory structures and motor programming units involved in speech sequencing remain unclear. Our results are consistent with theoretical accounts that posit the consonant cluster as a fundamental unit of phonological working memory and speech sequencing as well as those positing the syllable as a fundamental unit of motor programming.

Declaration of Competing Interest

None.

Acknowledgements

The research reported here was supported by a grant from the National Institutes of Health (R01 DC007683; F.H. Guenther, PI). We are grateful to Barbara Holland, Tess Fairchild and Riccardo Falsini for assistance with subject recruitment, data collection and analysis. This work benefited from helpful discussions with, or comments from, Stefanie Shattuck-Hufnagel, Jason Bohland, Jason Tourville, Elaine Kearney, and members of the audience at the 16th biennial Laboratory Phonology Conference.

References

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition, 104*, 591–630.

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer [Computer program]*. Version 6.0.53, retrieved 26 May 2019 from http://www.praat.org/.

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience, 22*(7), 1504–1529.

Buchwald, A., Gagnon, B., & Miozzo, M. (2017). Identification and remediation of phonological and motor errors in acquired sound production impairment. *Journal of Speech, Language, and Hearing Research, 60*, 1726–1738.

Buchwald, A., Calhoun, H., Rimikis, S., Lowe, M. S., Wellner, R., & Edwards, D. J. (2019). Using tDCS to facilitate motor learning in speech production: The role of timing. *Cortex, 111*, 274–285.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics, 24*(2), 209–244.

Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition, 99*, 205–235.

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics, 34*, 104–137.

Davidson, L. (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics, 38*(2), 272–288.

Dupoux, E., Kakehi, K., Pallier, Y., Hirose, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1568–1578.

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. (2011). Where do illusory vowels comes from? *Journal of Memory and Language, 64*(3), 199–210.

Fowler, C. A., Treiman, R., & Gross, J. (1993). The structure of English syllables and polysyllables. *Journal of Memory and Language, 32*, 115–140.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain & Language, 96*, 280–301.

Guenther, F. H. (2016). *Neural control of speech.* Cambridge, MA: MIT Press.

Hallé, P. A., Segui, J., Frauenfelder, U., & Meunier, C. (1998). Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance, 24*(2), 592–608.

Hindson, B. A., & Byrne, B. (1997). The status of final consonant clusters in English syllables: Evidence from children. *Journal of Experimental Child Psychology, 64*(1), 119–136.

Kirk, C., & Demuth, K. (2005). Asymmetries in the acquisition of word-initial and word-final consonant clusters. *Journal of Child Language, 32*(4), 709–734.

Kozhevnikov, V. A., & Chistovich, L. A. (1965). *Speech: Articulation and perception.* Washington, DC: Joint Publication Research Service.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50*, 239–269.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences. 22*, 1–75.

Loevenbruck, H., Collins, M. J., Beckman, M. E., Krishnamurth, A. K., & Ahalt, S. C. (1999). Temporal coordination of articulatory gestures in consonant clusters and sequences of consonants. In O. Fujimura, B. D. Joseph, & B. Palek (Eds.). *Proceedings of linguistics phonetics* (pp. 547–573). Prague: The Karolinum Press.

MacKay, D. G. (1970). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology, 3*(2), 210–227.

Pitt, M. A. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & Psychophysics, 60*(6), 941–951.

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., ... Eickhoff, S. B. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage, 60*, 830–846.

Segawa, J., Tourville, J. A., Beal, D. S., & Guenther, F. H. (2015). The neural correlates of speech motor sequence learning. *Journal of Cognitive Neuroscience, 27*(4), 819–831.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. F. MacNeilage (Ed.). *The production of speech* (pp. 109–136). New York: Springer.

Treiman, R. (1984). On the status of final consonant clusters in English syllables. *Journal of Verbal Learning and Verbal Behavior, 23*(3), 343–356.

Wilson, C., Davidson, L., & Martin, S. (2014). Effects of acoustic-phonetic detail on cross-language speech production. *Journal of Memory and Language, 77*, 1–24.