Journal of Phonetics 40 (2012) 374–389

Contents lists available at SciVerse ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Bridging planning and execution: Temporal planning of syllables

Christine Mooshammer [a,*], Louis Goldstein [a,b], Hosung Nam [a], Scott McClure [a,c], Elliot Saltzman [a,d], Mark Tiede [a]

[a] Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511, USA
[b] Department of Linguistics, Grace Ford Salvatori 301, University of Southern California, Los Angeles, CA 90089-1693, USA
[c] Nuance Communications, Inc., One Wayside Rd., Burlington, MA 01803, USA
[d] Department of Physical Therapy & Athletic Training, Boston University, 635 Commonwealth Ave., Boston, MA 02215, USA

ARTICLE INFO

Article history:
Received 23 September 2010
Received in revised form
7 February 2012
Accepted 9 February 2012
Available online 29 March 2012

ABSTRACT

This study compares the time to initiate words with varying syllable structures (V, VC, CV, CVC, CCV, CCVC). In order to test the hypothesis that different syllable structures require different amounts of time to prepare their temporal controls, or plans, two delayed naming experiments were carried out. In the first of these the initiation time was determined from acoustic recordings. The results confirmed the hypothesis but also showed an interaction with the initial segment (i.e., vowel-initial words were initiated later than words beginning with consonants, but this difference was much smaller for words starting stops compared to /l/ or /s/). Adding a coda did not affect the initiation time. In order to rule out effects of segment-specific articulatory to acoustic interval differences, a second experiment was performed in which speech movements of the tongue, the jaw and the lips were recorded by means of electromagnetic articulography. Results from initiation time, based on articulatory measurements, showed a significant syllable structure effect with VC words being initiated significantly later than CV(C) words. Only minor effects of the initial segment were found. These results can be partly explained by the amount of accumulated experience a speaker has in coordinating the relevant gesture combinations and triggering them appropriately in time.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

The act of producing an utterance results in vocal tract activity that is systematically patterned in time. This pattern somehow embodies the informational and contextual properties of the utterance in ways that are not yet completely understood. Investigation of this problem has largely been divided into two separate research areas: studies of internal *planning* processes (and their timing) that result in a sequence of linguistic units at several levels (phrases, words, feet, syllables, segments), that are ready to be executed, and studies of the *articulatory patterning* (in space and over time) associated with particular utterance types, observed as they are executed. While this implied distinction between planning and execution seems sensible enough, some recent work has suggested that we cannot treat them as entirely separate, modular activities (e.g., Goldrick & Blumstein, 2006; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007b; Lukatela & Turvey, 2000; Saltzman, Nam, Krivokapić, & Goldstein, 2008), though it is beyond the scope of this paper to visit these arguments. What is relevant here is to point out (based on evidence reviewed below) that some kind of control structure(s) that can guide relative timing of articulatory events must be in place in advance of the actual initiation of those events in the vocal tract. For the lack of a better term, we will refer to specification of temporal control in advance of articulation as a kind of "plan," and the deployment of a given control structure as a kind of "planning," without implying a commitment to the view that planning and execution are in fact distinct systems. Crucially, this is an aspect of planning that has not received attention in the planning literature, and limited investigation in the articulatory patterning literature.

Temporal plans for speech: Research on articulatory control of speech has revealed that speech can be decomposed into gestures, primitive compositional units of action (Saltzman & Munhall, 1989) and information (Browman & Goldstein, 1989) that form units such as segments, syllables and words, and which are arranged in time in a complex, partially overlapping fashion, rather than as a strict sequence of synchronous "bundles" (Browman & Goldstein, 1986, 1990b; Byrd & Tan, 1996; Fujimura, 1981, 1986; Mooshammer & Fuchs, 2002; Zsiga, 1995). The gestures that compose a single segment may in fact be produced sequentially, and those that compose successive segments may be produced

* Corresponding author. Tel.: +1 203 8656163x315; fax: +1 203 8658963.
E-mail address: tine.mooshammer@haskins.yale.edu (C. Mooshammer).

0095-4470/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2012.02.002

synchronously (see e.g., Bell-Berti & Harris, 1981; Fowler, 1980; Mooshammer, Hoole, & Geumann, 2006). Regularities in patterning have nonetheless been uncovered, and two examples are given here. First, a nasal consonant such as /n/ is composed of an oral constriction gesture (tongue tip for /n/) and a velum lowering (nasalization) gesture. In English, the onsets of these gestures have been shown to be synchronous when the nasal is in syllable onset (e.g., "pay nova"), but they are organized sequentially, with the velum gesture leading, when in coda (e.g., "pain over"; see Byrd, Tobin, Bresch, & Narayanan, 2009; Krakow, 1993). Second, consider the timing between a consonant oral constriction gesture in a syllable onset and the tongue gesture for the vowel. When there is a single consonant in the onset (e.g., "pot"), its constriction gesture onset is approximately synchronous for that of the following vowel (Löfqvist & Gracco, 2002), but in the case of an onset cluster (e.g., "spot"), the vowel gesture is not initiated synchronously with either consonant, but rather its onset is roughly at the midpoint of the consonant sequence (so-called "C-center" effect; Browman & Goldstein, 1988, 2000; Goldstein, Nam, Saltzman, & Chitoran, 2009; Kühnert, Hoole, & Mooshammer, 2006; Marin & Pouplier, 2010).

Regularities of timing in gesture initiation exhibit two hallmark properties which are relevant to temporal planning. One is that these regularities are language-dependent, which means that they do not fall out of very general neural or biomechanical properties. So if two events, A and B, are systematically produced with different relative timing in two languages, there must be some different control over relative timing (either direct or indirect) used by speakers of the two languages that must be in place internally before event B is initiated. We take this to be a kind of planning. The second is that the regularities crucially refer to syllable structure, therefore implicating syllable organization in the control structure responsible for regulating sub-segmental timing.

Language dependence can be demonstrated in both of the examples presented above. The timing of oral constriction and velum gestures in some languages, e.g. French (Cohn, 1993) and Arrente (Tabain, Breen, & Butcher, 2004), are synchronous in coda, as well as in onset, unlike the asymmetry observed in English. The C-center effect has been observed in word-initial consonant sequences in English (Goldstein et al., 2009; Marin & Pouplier, 2010), French (Kühnert et al., 2006), Italian (Hermes, Grice, Mücke, & Niemann, 2008), Mandarin (Gao, 2008), and Romanian (Marin, 2011). However, in Italian (Hermes et al., 2008), while stop-liquid onset clusters show the C-center effect, /s/-stop clusters fail to show it, unlike English which exhibits the effect for both cluster types. The C-center effect is generally not observed at all in Tashlihiyt Berber (Goldstein, Chitoran, & Selkirk, 2007a; Hermes, Ridouane, Mücke, & Grice, 2011) or Moroccan Arabic (Shaw, Gafos, Hoole, & Zeroual, 2009).

These examples also illustrate the central role of syllable structure, transparently so in the case of velum-oral coordination. In the case of C-center timing, the languages noted above that fail to exhibit the onset C-center effect (Berber, Moroccan Arabic) have been described as allowing only a single consonant in an onset, so the word-initial consonant sequences in these languages have a distinct syllable structure (e.g., the initial consonant is "extra-syllabic"). Also, while onset clusters exhibit the C-center effect in a variety of languages, coda clusters do not show the mirror image effect with anything like the same regularity (Byrd, 1995; Honorof & Browman, 1995; Marin & Pouplier, 2010).

Planning models: There are several influential models of speech production planning (Bohland, Bullock, & Guenther, 2009; Dell, 1986; Levelt, Roelofs, & Meyer, 1999; Vousden, Brown, & Harley, 2000) that operate in quite different ways, but roughly share the assumptions that important stages include accessing phonological information, constructing full phonological representations ("phonological encoding"), and controlling segmental sequencing. They generally do not address the issue of how the output sequence of phonological segments can be used to trigger the subsegmental gestural events at the appropriate times to yield the kind of regularities described above.

Two of these models that have paid some attention to this issue, however, are the GODIVA model (Bohland et al., 2009) and the WEAVER++ model (Levelt et al., 1999). In the GODIVA model, the segmental sequence that results from the planning process serves as input to the DIVA model (Guenther, 1995), that, in turn, produces patterns of articulator movement that achieve the sequence of (auditory and somatosensory) goals of the segment sequence. While this model does reproduce some observed patterns of coarticulation and could presumably learn the syllable-structure dependence of coarticulatory effects such as the velum-oral coordination described above (assuming that syllable-position information is also available in the input to DIVA), DIVA to date has not addressed temporal regularities on the scale of the C-center effects. In the WEAVER++ model, the frequent syllables of a language are stored in a "mental syllabary" (Levelt, 1992; Levelt & Wheeldon, 1994). The entries contain abstract gestural scores (which presumably contain some of the detailed relative timing information discussed above) stored as a whole and are retrieved during phonetic encoding. As they point out, however, these scores would have to abstract away from real time, as the same syllable can be produced at different rates and in different prosodic contexts, which will alter the actual time between events.

What is required is a control pattern (or plan) for inter-gestural timing that is flexible enough to be modulated as a function of these variables, while at the same time stable enough to express the gestural patterning regularities of a particular language. They do not attempt to further specify the details of such a control system, although the "gestural coupling graph" (Goldstein, Byrd, & Saltzman, 2006; Nam, 2007; Nam & Saltzman, 2003; Saltzman et al., 2008) described below is intended to be a representation at the appropriate level of abstraction. For infrequent syllables, the abstract gestural score (or coupling graph) needs to be assembled, but again, the model has nothing to say about the nature of this assembly process. The model does predict that high frequency syllables should be ready for execution more quickly than low frequency ones, and this effect has indeed been demonstrated in a task that eliminates lexical access effects (Cholin, Levelt, & Schiller, 2006). Relatedly, CVC words with high phonotactic probability are initiated more quickly in a picture-naming task than words with low phonotactic probability (Vitevitch, Armbruester, & Chu, 2004), when controlling for word frequency and neighborhood density. This could be interpreted as indicating that the gestural timing plans for frequently produced (high probability) sequences are well-learned and are either stored in memory or can be more quickly assembled. In sum, apart from the effects of learning/practice very little is understood about the assembly of motor (gestural) plans for an utterance.

Planning and syllable structure: The important role played by syllable structure in the plans for gesture timing suggests that probing the planning of different syllable structure types using reaction time methods would provide some insight into the nature of the temporal planning process. For example, the universality of CV syllables (Clements, 1990; Clements & Keyser, 1983) may indicate that their timing can be set up quickly and simply compared to syllables with coda consonants (CVC), clusters (CCV(C)), or those without onsets (V(C)). The experiments presented here are meant to be as a first step in this direction.

One model that would predict such syllable effects is the *coupled oscillator model* of timing planning (Goldstein et al., 2006;

Nam, 2007; Nam & Saltzman, 2003; Saltzman et al., 2008). In this model, coordination of gestures in time is achieved by associating each gesture with its own (internal) planning oscillator (or clock) that is responsible for triggering the activation of that gesture's control parameters in the vocal tract (which in turn results in the onset of motion of articulators controlled by that gesture). The clocks corresponding to the distinct gestures within a syllable are coupled to one another. The coupling relations are expressed in a *coupling graph*, in which the nodes are the gestural oscillators and the edges are coupling parameters governing that pair of oscillators: their coupling strength and (target) relative phase (see Saltzman & Byrd, 2000; Saltzman et al., 2008). Ensembles of coupled oscillators will exhibit entrainment: spontaneously settling into distinct, stable patterns of relative phase (Pikovsky, Rosenblum, & Kurths, 2001), where the particular pattern, or mode, is a function of the coupling parameters. In the coupled oscillator model, the planning oscillators are set into motion, at potentially arbitrary phases, and due to entrainment, they will settle into a pattern of stable relative phases. Once the phases stabilize, gestures will be triggered by their corresponding clocks.

The most basic coupling targets employed in the graphs correspond to coupling modes that have been shown to be intrinsically accessible (without training) to human actors when asked to rhythmically oscillate multiple body parts (see Haken, Kelso, & Bunz, 1985; Turvey, 1990)—*in-phase* and *anti-phase*. These two distinct modes have been hypothesized by Goldstein et al. (2006) to form the basis for syllable structure in human language: oscillators for onset consonant gestures are coupled in-phase with the (nuclear) vowel oscillator (so they are triggered synchronously), while coda consonant oscillators are coupled anti-phase with the vowel (and are triggered later than the vowel). Setting up the temporal plan for gestures that are coupled in-phase would be expected to be faster than for those coupled in anti-phase, either because of accessibility of the in-phase mode during assembly of the graph, or due to reduced time required for in-phase oscillators to settle. Consistent with this hypothesis, Nam (2007) conducted a pilot experiment testing the prediction that time to initiate production should be faster for CV syllables than for VC. Two speakers each of Korean and English were presented with random combinations of a C letter (voiceless stop consonants, "ptk") and a V letter ("aiu") in English that appeared on a screen arranged vertically. They were instructed to assemble syllables from these letters, beginning with the letter on the top of the screen, and to produce the syllable as quickly as possible. He found that CV syllables were initiated significantly faster than VC as measured acoustically from the reaction time.

Design considerations: A general problem with using standard reaction time methods to study syllable structure is that word forms such as CV vs. VC will necessarily begin with different segments. Reaction time has generally been measured from the acoustic signal and depending on the nature of the initial gesture to be produced, the time from articulatory initiation (which is what we are interested in from the perspective of speech *production*) to the acoustic onset (the so-called articulatory to acoustic interval, or AAI, see Kawamoto, Lui, Mura, & Sanchez, 2008) can be quite variable (see Kessler, Treiman, & Mullennix, 2002; Rastle, Croot, Harrington, & Coltheart, 2005). Vowels would be expected to exhibit acoustic consequences much earlier in their articulation than would be the case for voiceless stop consonants, that is, vowels are *expected* to have a *shorter* AAI than stop consonants. Because it is impractical to run a large group of subjects in an articulatory kinematic experiment, we decided to undertake two versions of a naming task; audio-only (Experiment 1), and audio recorded concurrently with articulatory motion (Experiment 2). Comparing the results of these two experiments is important not only for beginning to reveal differences in planning time as a function of syllable structure, but also is relevant to interpreting many results in the literature that have relied on acoustic criteria alone for vocal response initiation.

Another issue in the design of an experiment probing the assembly of gestural timing plans is that potentially confounding effects due to lexical access and retrieval of phonological information, for example word frequency or neighborhood density, should be avoided. To avoid these confounds, we decided to employ a delayed naming paradigm (e.g. Rastle et al., 2005), which several studies have shown are free from frequency effects (e.g. Goldinger, Tamiko, Abramson, & Jain, 1997; Monsell, 1990) and neighborhood density effects (e.g. Grainger, Spinelli, & Ferrand, 2000; Mulatti, Reynolds, & Besner, 2006). In this kind of task speakers are asked to produce already selected syllables as quickly as possible after a GO signal. It has been used to investigate speech execution as isolated from lexical retrieval and phonological encoding. In some cases it has been used as a control condition in combination with simple naming to factor out "execution-level" effects, so that "pre-execution" processes can be investigated (e.g. Grainger et al., 2000). The logic is that effects that show up in simple naming, but not in delayed naming, are assumed to result from "pre-execution" retrieval and planning. In other work using delayed naming, the nature of the execution processes is itself the object of investigation (Monsell, 1990; Sternberg, Monsell, Knoll, & Wright, 1978). In general, the assumption has been that in delayed naming of single, syllable-sized units, the time to initiate production, as measured acoustically, depends only on the nature of the initial segment and to a lesser degree the second segment (Kessler et al., 2002). (Kessler et al. (2002) and Rastle et al. (2005) showed RT measured acoustically varied across initial consonant types, because the AAI depends on the aerodynamic and acoustic nature of the consonant.) However, the processes that we are probing, namely, the planning of intergestural timing for the syllable, lie at the margin of what are generally considered planning and execution. Because planning at the level of temporal control is so bound up with the ongoing articulation, it is reasonable to expect that these processes are taking place concurrently and so might be expected to affect delayed naming reaction time.

While the effects of factors such as word frequency and neighborhood density can be expected to be mitigated by the use of delayed naming, factors such as syllable frequency and phonotactic probability should reflect the degree to which a gestural pattern is "well practiced," and can thus be expected to influence the time required to assemble or deploy a gestural timing plan, and thus affect delayed naming reaction time. Indeed, the design of the experiment by Cholin et al. (2006) involved learning the association between four similar items and positions on a screen. This reduced lexical and phonological encoding effects to a minimum, and syllable frequency effects were nonetheless obtained. There is no reason to expect that such effects are limited to syllables, however, as there is also evidence for pre-assembled motor plans for units larger than syllables, such as words or phrases (Kapatsinski, 2010; Pierrehumbert, 2003), and phonotactic probability effects could indicate smaller units as well. Unfortunately, given the relative markedness of different syllable structures, structure will necessarily be conflated with syllable frequency and phonotactic probability, particularly in a design that requires a relatively small number of minimal pair stimuli (that is, segmental composition needs to be the same across CV and VC stimuli, for example). The relative contribution of structure and probability variables in explaining any main effects will have to be evaluated statistically post hoc.

Finally, Kawamoto et al. (2008) have recently shown that the delayed naming reaction time also depends on *task* variables, such as the length of the delay, and how subjects are instructed

with respect to maintaining their articulatory postures during the delay interval. With long delays and "standard" instructions, which do not control the subjects' articulation during the delay interval, subjects can effectively begin execution during the delay interval, by completing much of the articulation of the first segment during the delay interval but withholding phonation until the response signal. Thus, it is also important to control the subjects' articulation during the delay interval, which is accomplished in the experiments reported here by requiring the subject to produce a schwa vowel ('uhhh') from the time the stimulus is presented until they receive the GO signal.

2. Experiment 1: measuring planning time acoustically

2.1. Method

2.1.1. Participants

Twenty-two native speakers of American English, most of them college students from the Yale University community, participated in this experiment. They were paid for their participation. Due to technical problems two female speakers had to be excluded. Thus, data of 12 female and eight male speakers were used. No participant reported any history of speech or hearing impairment.

2.1.2. Materials

The material consisted of 45 words and nine non-words with varying syllable structure (V, VC, CV, CVC, CCV, CCVC), initial consonants (/p, t, k, s, l/) and vowels (/i, ei/) (see Table 1). Only words with tense vowels were chosen here because lax vowels do not occur in open syllables. For tense vowels this restriction does not exist. Stimulus sets were grouped according to final consonants for VC words, e.g. /t/ in *eat*, and initial consonants for C(C)V(C) sequences, e.g. *tea*. Since for some combinations of vowels and consonants, no real words were available, nonsense sequences following English phonotactic rules were used instead, e.g. *eap* or *tay*. Despite the known disadvantages, words and non-words were mixed here. Since both vowels and consonants could be initial segments, the selection of the segments used for composing the CV and VC syllables had to be very closely controlled. These strong restrictions on syllable structure and segmental identity also meant that well-known influences on lexical access, such as word frequency, familiarity or neighborhood density, could not be controlled in the selection of the material. In order to deal with these effects, reaction times were measured in a delayed naming paradigm (see Section 2.1.3 for details).

Table 1
Stimuli included in Experiments 1 and 2. Items in brackets were not included in the corpus for Experiment 2. Non-words are printed in italics.

Consonant	CCVC	CCV	CVC	CV	VC	V
No C						(A I)
Lateral			lake (leak)	lay (lee)	ale (eel)	
Stops			*pape* (pate) (peep peak) tape (take) (teet teak) cape (cake) (*keek* keep)	pay pea *tay* tea Kay key	ape *eap* ate eat ache eke	
Fricative	(spate state) (skate slate) (speak steep) (skeet sleek)	(spay stay) (*skay* slay) (*spee stee*) (ski *slee*)	sake (seat)	say (sea)	ace (*eese*)	

In Table 2 lexical characteristics of the items used for this experiment are presented grouped by syllable structure. The following lexical and structural properties of the test items were included: word frequency, neighborhood density, onset density, phone and biphone probability, and syllable frequency. The first two rows give the total number of items and the number of non-words.

All figures in this table that pertain to corpus frequency are drawn from the CELEX Corpus (Baayen, Piepenbrock, & Rijn, 1995), based on the 17.9 million words of the COBUILD corpus. The means of word frequencies, given in the third line, were calculated by adding the number of tokens appearing in this corpus to the number of tokens of regular inflections of these tokens. Without adding the tokens of inflections, some words (e.g. *pea*: 30) have a very low number of occurrences, even though they are very frequent in inflected forms (e.g. *peas*: 146). The words *Kay* and *eke* did not occur in the corpus. In order to avoid too many missing values and to distinguish rare words from non-words, 1 was added to the occurrences of all real words and then the logarithm was calculated. As can be seen, on average, words with CV structure had a higher word frequency than words with VC structure. However, these differences did not reach significance in pairwise t tests. Neighborhood density calculations were based on the database provided by the Washington University Speech and Hearing Laboratory. The number of phonological neighbors for each word is determined by finding, for each word, the number of words in the database whose familiarity rating is 6 or higher and that match the phonological form of the target word exactly, except for the insertion of one phone, the deletion of one phone, or swapping the order of two phones. Denser neighborhoods have been found to facilitate lexical retrieval and therefore items with many neighbors have shorter reaction times. As can be seen in Table 2 items starting with clusters have come from sparser neighborhoods than V(C) items, which in turn have significantly fewer neighbors than CV items. The existence of a coda does not significantly affect this value. Onset density, given in the next row, refers to the proportion of neighbors that share the same onset. Vitevitch et al. (2004) found that words with sparser onset neighborhoods are initiated quicker and produced more accurately than words with denser onset neighborhoods.

Table 2
Characteristics of items used in Experiment 1. The reported significant differences are based on pairwise *t*-tests with a Bonferroni adjustment for multiple testing.

Characteristics	V	VC	CV	CVC	CCV	CCVC	Sig. diff.
Total number	2	10	10	16	8	8	
Non-words		2	1	2	4		
Word frequency	3.8	2.5	4.6	2.9	4.7	2.9	n.s.
Neighborhood density	24.5	20.4	26.4	22.8	10.9	9.5	CCV(C) < (C)V(C)VC < CV
Onset density	7.0	7.4	11.0	12.9	6.1	6.1	V(C), CCV(C) < CV(C)
Phone probability	0.004	0.015	0.0511	0.0517	0.0487	0.0557	V(C) < C(C)V(C)
Biphone probability	NA	0.0003	0.0019	0.0020	0.0056	0.0042	VC < CV(C) < CCVC < CCV
Syllable frequency	9.62	2.41	7.16	2.73	3.87	2.54	VC, CVC, CCV(C) < CV < V

In our data the onset was defined as the initial vowel in V(C) syllable, as the initial consonant in CV(C) words and as the initial cluster in CCV(C) words. In our corpus, CV(C) come from a denser onset neighborhood than either V(C) or CCV(C) words. The next row of Table 2 shows the mean phonotactic probabilities for the phones occurring in different syllable structures. Following Vitevitch and Luce (2004), two measures were used for phonotactic probability: phone probability and biphone probability. Phone probability was calculated as the sum of the left-to-right position-specific probabilities of the phones that are found in the word, divided by the number of phones in the word. In other words, this number gives the average probability of the phones that appear in the word in a given position. Vowel initial items had a significantly lower mean phonotactic probability than items starting with consonant for the items used in this study. Biphone probability additionally increases significantly for cluster initial words. Items with a higher phonotactic probability were found to be initiated quicker and with less frequent errors than items with a lower phonotactic probability (see Vitevitch et al., 2004). Following Levelt and Wheeldon (1994) syllable frequency was calculated based on the CELEX database as the log frequency of occurrences of each phonetic syllable in the corpus. In general, VC and CVC words had a significantly lower syllable frequency than CV words.

2.1.3. Procedure and apparatus

In order to control for word frequency effects we employed the delayed naming paradigm (see Kawamoto et al., 2008; Rastle et al., 2005; Sternberg et al., 1988). This was implemented using custom software that controlled presentation of each stimulus on a computer screen as follows: first, subjects were shown the target word with a "Get ready, say 'uhhh'" prompt, their cue to read silently, inhale, and produce schwa. (Schwa was elicited in order to avoid articulatory preparation before the imperative signal; see Kawamoto et al., 2008 for a detailed discussion of preparatory movements in delayed naming tasks.) Next, at a randomized delay varying between 1 and 2 s, the screen changed color, the prompt changed to "Go," and an audible beep was emitted, providing the cue to the subject to produce the target word as quickly as possible. Individual trials were separated by fixed 250 ms intervals.

When each subject arrived, they received written instructions about the task. They were seated in a quiet room with the computer display in front of them. First, they were familiarized with the material used here by going through the whole list twice, in order to ensure that the speakers pronounced all non-words and words correctly. Then they were trained to start uttering the schwa after the "Get ready, say 'uhhh'" message and the test item appeared on the screen. The subjects were corrected by the experimenter during this training session if they produced the test item before the cue signal. The task was repeated about 10 times during the training session.

The 54 stimulus items were repeated six times in a block-wise randomized order. The experiment also included a second experimental condition without the preproduction schwa ("simple condition"). Subjects were randomly assigned to start with either condition. Since the results of the prevocalic and the simple condition are very similar apart from much longer reaction times in the magnitude of 200 ms in the condition without schwa, the results of the simple condition are not reported here.

2.1.4. Measurements

The stimulus onset was determined by automatic detection of the maximal amplitude of the beep. This signal was a sinusoid of 500 Hz, amplitude filtered using a triangular window with a duration of 90 ms. The onset of the speech signal was labeled manually at the release burst for items beginning with stops, at the onset of frication for initial fricatives, and at the end of the transition from the preceding schwa to a lateral. For vowel initial sequences, most speakers either interrupted voicing between the schwa and the following vowel (see upper panel in Fig. 2) or they glottalized. Occasionally, they also produced short transitions between the schwa and the vowel. The onset of the vowel was labeled as the first pulse with the visible formant structure. Sometimes this time-point occurred during irregular voice periods. The interval from the peak of the beep to the acoustic onset of the initial segment was then used as an acoustic measure for reaction time, RT_{ac} for short.

False or uncued early starts were excluded by omitting all trials with a negative reaction time or when the reaction time was smaller than 50 ms, which was the case in 4.4% of all 7521 trials. Since the time of the beep was measured at the peak and not the onset (a difference of 45 ms), the exclusion criterion is somewhat shorter than in other studies. Further outlier detection was based on the residuals of the Linear Mixed Effects Models as suggested by Baayen (2008): items with residuals exceeding 2.5 times of the standard deviation were considered outliers and therefore removed. These numbers will be reported in the corresponding results section.

2.1.5. Statistics

The hypothesis that reaction times are shorter for words with C(C)V(C) structure than words with V(C) structure was tested by using Linear Mixed Effects Models (see e.g. Baayen, 2008; Pinheiro & Bates, 2000, & others) with RT_{ac} as the dependent variable, and subject as a random factor. Fixed effect factors included were: syllable structure STRUCT with the levels V, VC, CV, CVC for the sets with stops and the lateral, and additionally CCV and CCVC for the fricative set, item vowel VOWEL with levels /i/ and /ei/. Interactions were included in the model. The data were grouped into three sets according to the type of consonant (stop, lateral, or fricative) that appeared in the onset of C(C)V(C) stimuli and in the coda of VC stimuli. This grouping was necessary because syllable structure was not fully crossed with consonant type (i.e. there were no initial clusters for the /l/ and the stop set) and because the interaction between STRUC and CONS was significant. The words consisting of a single vowel or diphthong (STRUCT = V) had to be reused in all three sets.

In order to avoid collinearity between factors and factor levels, the factors were coded and centered by subtracting the grand mean, following the suggestions by Gelman and Hill (2007). All statistics were carried out using R 2.8.1 (see R 2008). In a first step it was tested whether the fit improved by allowing the slopes of the fixed effects to vary across subjects using a log-likelihood test for goodness of fit. If the inclusion of a by-factor random slope improves the model significantly, it means that the subjects do not behave consistently regarding levels of this factor. For significant interactions the dataset was split accordingly. Statistical results of the fixed effects are presented by the estimates of the regression coefficients of the model β and the standard error of β. Instead of ANOVA tables with F values, p values are presented based on Markov Chain Monte Carlo samples with 1000 simulations. This method is generally preferred because the correct calculation of degrees of freedom is still controversial for Linear Mixed Effects Models. Planned comparisons were carried out for significant multilevel factors, such as STRUCT, based on models with uncentered factors and p adjustments for multiple comparisons using the Holm algorithm.

2.2. Results

Fig. 1 shows the means of RT_{ac} sequences split by the consonant types stops, /l/ and /s/. For each consonant type the

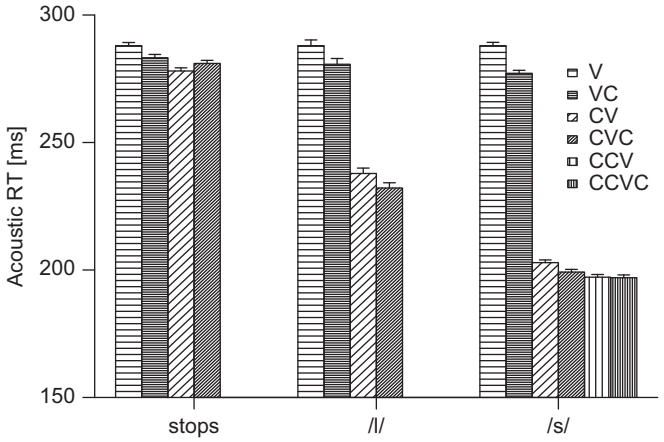

Fig. 1. Means and standard errors of the acoustically measured reaction-time RT_{ac} for different syllable structures, grouped by consonant types. For each consonant type the subset consists of VC syllables with identical coda consonants and of C(C)V(C) syllables with identical initial onset consonants. Words with V structure were reused for all three sets.

Table 3
Linear Mixed Effects Model for RT_{ac} split by consonant groups, and the results from pairwise *t*-test comparisons.

Consonant	Effect	β	SE(β)	*t*-value	p_{MCMC}	Sig. diff.
Stops	Intercept	280.744	9.60			
	STRUCT	−2.01	2.24	−0.90		
	VOWEL	−5.3	1.97	−2.69	**	/i/ > /ei/
	STRUCT × VOWEL	−1.88	1.97	−0.95		
Lateral	Intercept	259.18	8.85			
	STRUCT	−21.60	3.34	−6.48	***	V(C) > CV(C)
	VOWEL	3.46	3.57	0.97		
	STRUCT × VOWEL	−1.30	3.18	−0.41		
Fricative	Intercept	210.97	6.96			
	STRUCT	−17.19	1.41	−12.23	***	V(C) > CV(C), CCV(C)
	VOWEL	−0.72	1.98	−0.38		
	STRUCT × VOWEL	0.81	1.23	0.66		

'.': $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

data subset consisted of VC syllables with coda consonants identical to the onset consonants of the C(C)V(C) syllables in that group. This kind of grouping was necessary because the factors STRUCT and CONS could not be fully crossed and because the interaction between the two factors was highly significant (when excluding the clusters). Items with STRUCT = V were reused for each set. The standard errors are based on the cell means, i.e., standard deviations were divided by the square root of the number of speakers times the number of items. Out of 6715 items, 182 were excluded as outliers. The median RT_{ac}s for the onset-less stimuli (V(C)) are longer than the RT_{ac}s for stimuli with onsets.

Table 3 gives the results of Linear Mixed Effects Models for RT_{ac} for each consonant set. For the stops, the model improved by including a speaker specific slope for the factor STRUCT. STRUCT did not affect RT_{ac}, but VOWEL did with longer reaction times for words with the vowel /i/. The difference between V(C) and CV(C) of 4 ms was not significant but in the hypothesized direction. For the data set with lateral consonants inclusion of a speaker-specific slope for the factor STRUCT improved the model significantly. Planned comparison showed that adding a coda did not have a significant effect on RT_{ac}. The significant main effect STRUCT corresponds to an average increase in RT_{ac} of 50 ms from V(C) to CV(C). There was no significant main effect of VOWEL or interaction between VOWEL and STRUCT. As can be seen from Fig. 1 the strongest effect of syllable structure on RT_{ac} was found for items with fricatives.

Table 4
Results from Experiment 1: Linear Mixed Effects Model for the acoustically measured reaction time RT_{ac} for all effects split by structure with significant pairwise *t*-tests comparisons.

Subset	Effect	β	SE(β)	*t*-value	p_{MCMC}	Sig. diff
CV(C)	Intercept	259.55	9.01			
	CONS	−39.92	2.72	−14.66	***	stops > /l/ > /s/
	VOWEL	−4.12	1.89	−2.18	*	/i/ > /ei/
	CONS × VOWEL	3.20	2.52	1.23		
VC	Intercept	281.38	10.04			
	CONS	−1.95	2.40	−0.81		
	VOWEL	−3.43	3.24	−1.06		
	CONS × VOWEL	−0.55	4.16	0.13		

'.': $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

The final model included a speaker-specific slope for the factor STRUCT. Planned comparisons showed that the difference of 81 ms between V(C) and CV(C) was highly significant. Furthermore, syllables starting with single consonants (CV(C)) had a marginally significantly longer RT of 4 ms than syllables starting with clusters. To summarize so far, onsetless syllables take longer to initiate than syllables starting with a consonant (though not significantly so for stops). Adding a coda did not affect the RT in our data. Adding another consonant in the onset did reduce the RT slightly for fricative initial words.

As can also be seen in Fig. 1 the identity of the consonant affected the reaction time only for CV(C) words, but not for V(C) words. In order to test this, the data were split by structure (Table 4). Since adding a coda did not affect RT_{ac}, CV and CVC words were pooled. No clusters were included because all clusters started with a fricative. Inclusion of a speaker-specific slope did improve the model significantly. The identity of the consonant had a significant effect for CV(C) sequences with stops exhibiting the longest RT_{ac}, followed by the lateral. Items with an initial /s/ were produced with the shortest RT_{ac}. This is in agreement with results from the literature, see e.g. Kessler et al. (2002), Rastle et al. (2005), and Kawamoto et al. (2008). A significant main effect of VOWEL was found: words containing the vowel /i/ were initiated 4 ms later than words with the diphthong /ei/. As was expected, no significant consonant effects could be found for the VC sequences.

2.3. Discussion

In Experiment 1, the hypothesis that syllable structure affected initiation time was supported. CV(C) sequences were initiated more quickly than V(C) sequences for all consonant types, and this was statistically confirmed for the lateral and fricative conditions. Adding a coda did not affect the measured reaction time. Words with initial clusters also showed a small RT advantage over words beginning with a single C.

Role of lexical variables: While the observed differences in RT could reflect intrinsic differences between V(C) and CV(C) syllable structure types in the time required to prepare a gesture timing plan, they could also be due to one or more of the lexical variables presented in Table 2, all of which (except word frequency) differed significantly across syllable structure conditions. To test the potential role of these variables in accounting for the results, regressions were performed for each lexical variable between the measure's value for each stimulus item and the RT for that stimulus item (averaged across subjects). This was done over all stimuli and also separately for the stop, lateral and fricative subsets. The results are shown in Table 5. The first three variables shown (word frequency, neighborhood density, and onset density) show little or no ability to predict the RT values, either

Table 5
Slopes and adjusted R^2 for acoustically measured reaction time RT_{ac} and several word-based measures. Data are averaged across speakers. Values in column *All* include data for *V*. Values in italics show significant effects against predictions from the literature.

Characteristics	Variable	All	Stops	Fricative	Lateral
Word frequency	Slope	−1.3	0.2	−1.1	−12.0 .
	R^2	−0.02	−0.05	−0.05	0.45
Neighborhood density	Slope	*2.5***	−0.6	0.7	−0.6
	R^2	0.17	−0.11	0.05	−0.24
Onset density	Slope	1.1	−1.0*	−0.1	−3.8
	R^2	0.01	0.21	−0.05	0.31
Phone probability	Slope	−1237***	−212**	−1677***	−3250*
	R^2	0.28	0.25	0.73	0.64
Biphone probability	Slope	−14,555***	−5790***	−6613*	−20,973*
	R^2	0.51	0.40	0.22	0.59
Syllable frequency	Slope	0.4	−1.0	−2.1	−4.4
	R^2	−0.02	0.08	−0.03	−0.03

'.': $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

overall or in the subsets (neighborhood density shows a significant regression, but only in the overall analysis). While these variables are known to affect naming time in production experiments (Levelt et al., 1999; Oldfield & Wingfield, 1965; Vitevitch, 2002), the effects have been attributed to lexical access effects (Harley, 1993; Jescheniak & Levelt, 1994; Levelt et al., 1999) or the time required to activate the segments of the word in interactive activation models (Dell & O'Seaghdha, 1992; Vitevitch, 2002). Such effects are presumed here to occur without having necessarily assembled a complete coupling graph for production, based only on the gestural or segmental content. There has been no clear evidence for these effects in a delayed naming paradigm, in which there is adequate time to access the item and to activate its segments (see Goldinger et al., 1997; Grainger et al., 2000; Monsell, 1990; Mulatti et al., 2006). So it is not surprising that they are not predictive of RT in this experiment. Phonotactic probability (both phone and especially biphone) accounts for a large and significant proportion of the cross-item variance in RT. A higher probability is associated with faster RTs. While effects of these variables on naming time have also been attributed to the time required to activate the segments of the word (Vitevitch & Luce, 2004), their emergence in a delayed naming paradigm, combined with the failure of the other lexical variables to show substantial effects, suggests they reflect a different source in this experiment. Since the probability measures provide some estimate of a speaker's experience in actually producing these forms, the effect could arise from differential amounts of experience or practice with the particular gestural timing plan (or coupling graph) of these utterances.

By similar logic (and the experimental results of Cholin et al., 2006), we would expect that syllable frequency would also be predictive of RT, which is not the case. In fact, however, if what is relevant to this task is experience setting up the temporal coordination of the initial two segments (C–V or V–C), then the failure to obtain syllable frequency effects actually makes sense. For CVs with a tense vowel, the frequency of the CV syllable is generally going to be much higher than the CVC that includes that same CV. This is true for all the CV(C) examples in our stimuli. For example the frequency of *tea* is 1062, while that of *teak* is 23 and *teet* is 8. This is true for a couple of reasons: the CV sequences can occur in many more polysyllabic words than can the CVC, and also, the CV words tend to be higher frequency words than the CVCs (reflecting an overall preference for CV). But such differences in frequency are not going to be predictive of RT results, if it is only the temporal coordination of the initial two units that matters, as is suggested by the lack of any CV–CVC difference in RT. On the other hand, the phone probability measures do not distinguish CVs and CVCs in this way. If a particular CV sequence is common in English that will also contribute to the phone probability (and biphone probability) of the CVC that includes that same CV. So for example, *teak* actually has slightly higher phone and biphone probabilities than *tea*.

The faster reaction time for words beginning with /sC/ clusters (compared to initial Cs) could also be due either to their structural or lexical properties. The coupled oscillator planning model (see Nam, 2007) predicts faster oscillator settling time for an initial cluster than for a single C. It is also the case that the biphone probability of the clusters is significantly higher than for any other structural type. Rastle et al. (2005) also found shorter reaction times for initial clusters in a delayed naming task. Contrary to Kawamoto and Kello (1999)'s study they only found the cluster effect for stops and not for fricatives. They explain this asymmetry by hypothesizing a shortening of consonants within clusters. For the stop, shortening of the closure phase causes an earlier burst whereas for a fricative shortening affects the duration but not the acoustic onset. However, in the results presented here, only clusters with fricatives were investigated, and they did show an earlier acoustic onset.

The variables STRUCT, biphone probability and phone probability are highly correlated. In order to test which of the three variables might be more important in accounting for the RT differences, three LME models were calculated and the improvement of the model evaluated by χ^2 log-likelihood testing. The first model included only the factor STRUCT, the second only the phone and biphone probability variables, and the third one included all three variables. In this analysis, stimuli with onsetless open syllables (i.e. 'V') were excluded (because they have only one phone), and open and closed CV(C) and CCV(C) syllables were pooled because no effect was found for adding a coda (so there are three levels of STRUCT: VC, CV(C) and CCV(C)). Adding phone and biphone probability improves the model significantly ($\chi^2 = 48.2$, $p < 0.001$) compared to the model with only STRUCT. Likewise, adding STRUCT improves the model compared to the model with only phone and biphone probability ($\chi^2 = 392.0$, $p < 0.001$). Thus, both the phonotactic probability measures and syllable structure contribute to the observed reaction time differences in this experiment.

Role of measurement technique: The difference in delay between CV(C) and V(C) varied with the consonant involved. Sequences with stops showed only a small, non-significant difference. CV(C) sequences starting with the fricative /s/ were initiated much earlier than V(C) sequences. Laterals were approximately in-between stops and fricatives. Thus, within the

subset of CV(C) words, the reaction time was highly dependent on the initial consonant type. Shorter reaction times for fricatives and laterals as compared to stops have been found in several studies (e.g. Kessler et al., 2002; Rastle et al., 2005). This pattern can be simply explained by the fact that reaction time for stops was measured at the release burst, and therefore the RT_{ac} included the entire (silent) constriction duration. For fricatives, RT_{ac} was measured at the onset of fricative turbulence, which coincides with the very beginning of the constriction interval, so the constriction interval is *not* included in the RT in this case. For the lateral it was more difficult to detect the acoustic onset in a consistent manner. In general, therefore, acoustical measurements do not reflect the initiation of the gesture towards the first segment but only their acoustic consequences (see Kawamoto et al., 2008). In the standard interpretations of this task (see e.g. Sternberg et al., 1978, 1988) it is assumed that the interval from articulatory initiation to its acoustic signature in the sound stream (the AAI, see Introduction) is constant across different speech sounds, and therefore the articulatory onset can be derived from the acoustic measurements (see e.g. Sternberg et al., 1978, 1988). Since this is clearly not the case when comparing stops, fricatives, and vowels, the design was replicated in Experiment 2, while collecting articulatory kinematics.

A difference in the AAI potentially could also account for the longer latencies in VC than CV sequences, since vocalic gestures are known to have longer durations (time to reach their constriction targets) than consonantal ones (see e.g. Browman & Goldstein, 1990a; Perkell, 1969). The acoustic signature of initial vowels depends mainly on phonation onset and not on lingual articulation, while the acoustic signature of a fricative, for example, will occur as soon as the constriction is close to its target position. Consequently, if vocalic and consonantal gestures were initiated with the same latencies and if the phonation onset for the vowel is timed to coincide with its achievement of target, then the acoustic onset could be later for vocalic gestures due to their longer duration. This alternative explanation to the assembly hypothesis will also be tested with articulatory data in the next section.

3. Experiment 2: measuring planning time using articulatory data

3.1. Method

3.1.1. Participants

Four native speakers of American English (three females and one male), again college students from the Yale University community, participated in this experiment. They were paid for their participation and they also participated in the Experiment 1. Therefore training for the material or the procedure was not necessary.

3.1.2. Materials

The material was almost the same as in Experiment 1. All items were repeated eight times. Due to technical problems, only seven repetitions could be recorded for the male subject. In articulatory experiments more repetitions are necessary because trials might not be useable due to faulty sensors. For two speakers, the whole corpus was recorded in a "simple" condition without the sustained schwa during the delay interval, as well as the postvocalic condition with the sustained schwa. After inspecting the data, the simple condition was excluded because the speakers moved their articulators too much during the delay interval, which made it impossible to determine the movement onset for the stimulus production. Therefore, the last two speakers were only recorded in the postvocalic condition. In order to keep the whole experiment within a manageable time-frame of about 2 h, only for stops were CVs and VCs produced in both vowel contexts. Therefore, vowel identity is not a fully crossed condition in this experiment. Furthermore, we dropped the cluster and the single vowel condition. These changes in material affected the characteristics of word frequency, neighborhood density, onset density, phone and biphone probability and syllable frequency only slightly for words with CVC, CV and VC structures (see Table 6). Items with VC structure had significantly lower phone and biphone probabilities as well as lower syllable frequencies than CV and CVC words. Onset density was significantly lower for VC than for CVC words. Neighborhood density and word frequency did not differ significantly for varying syllable structures.

Table 6
Characteristics of items used in Experiment 2, articulatory data. The reported significant differences are based on pairwise *t*-tests with a Bonferroni adjustment for multiple testing.

Characteristics	VC	CV	CVC	Sig. diff.
Total number	8	8	5	
Non-words	1	1	1	
Word frequency	2.50	4.95	4.62	n.s.
Neighborhood density	20.75	24.87	20.60	n.s.
Onset density	7.7	9.9	12.6	VC < CVC
Phone probability	0.015	0.0513	0.0482	VC < CV(C)
Biphone probability	0.0003	0.0017	0.0019	VC < CV(C)
Syllable frequency	2.41	7.11	2.48	VC, CVC < CV

3.1.3. Procedure and apparatus

The same procedure as for Experiment 1 was used in Experiment 2, i.e. the delayed naming paradigm in a postvocalic condition. Additionally, speech movements of the tongue, the jaw and the lips were recorded using an ElectroMagnetic Midsagittal Articulometer system (EMMA) (Perkell et al., 1992). This system allows tracking of individual flesh-points in the midsagittal plane by attaching sensors to the articulators using dental adhesive. Alternating magnetic fields are generated by three transmitter sensors mounted on a helmet. When small sensors enter this field, alternating voltages are induced which are proportional to the distance to the transmitters. For the current experiment, three sensors were glued to the tongue. The frontmost sensor, T1, was placed about 1 cm behind the tongue tip. The rear-most sensor, T3, was positioned as far back as the subject would tolerate. T2 was placed in-between T1 and T3. One sensor on the lower incisors was used for monitoring jaw movements. Lip aperture was derived from calculating the Euclidean distance between sensors on the upper and lower lip. In order to compensate for head movements, the signals of two reference coils on the nasion and the gum above the upper incisors were used to translate and rotate the data to a consistent reference centered on the upper incisors.

3.1.4. Measurements

Measurements of RT_{ac} were carried out in the same way as in the acoustic-only experiment. The labeling procedure for the articulatory data is shown in the lower part of Fig. 2 for the item /pi/. The onset of the articulatory movement towards the initial segment was determined by using a 20% threshold of local peak velocity based on the velocity signal of the presumably most relevant sensor for the oral constriction gesture of that segment. For bilabials the lip aperture was used, and for the coronal consonants /t, l, s/ the onset was detected using the 2D

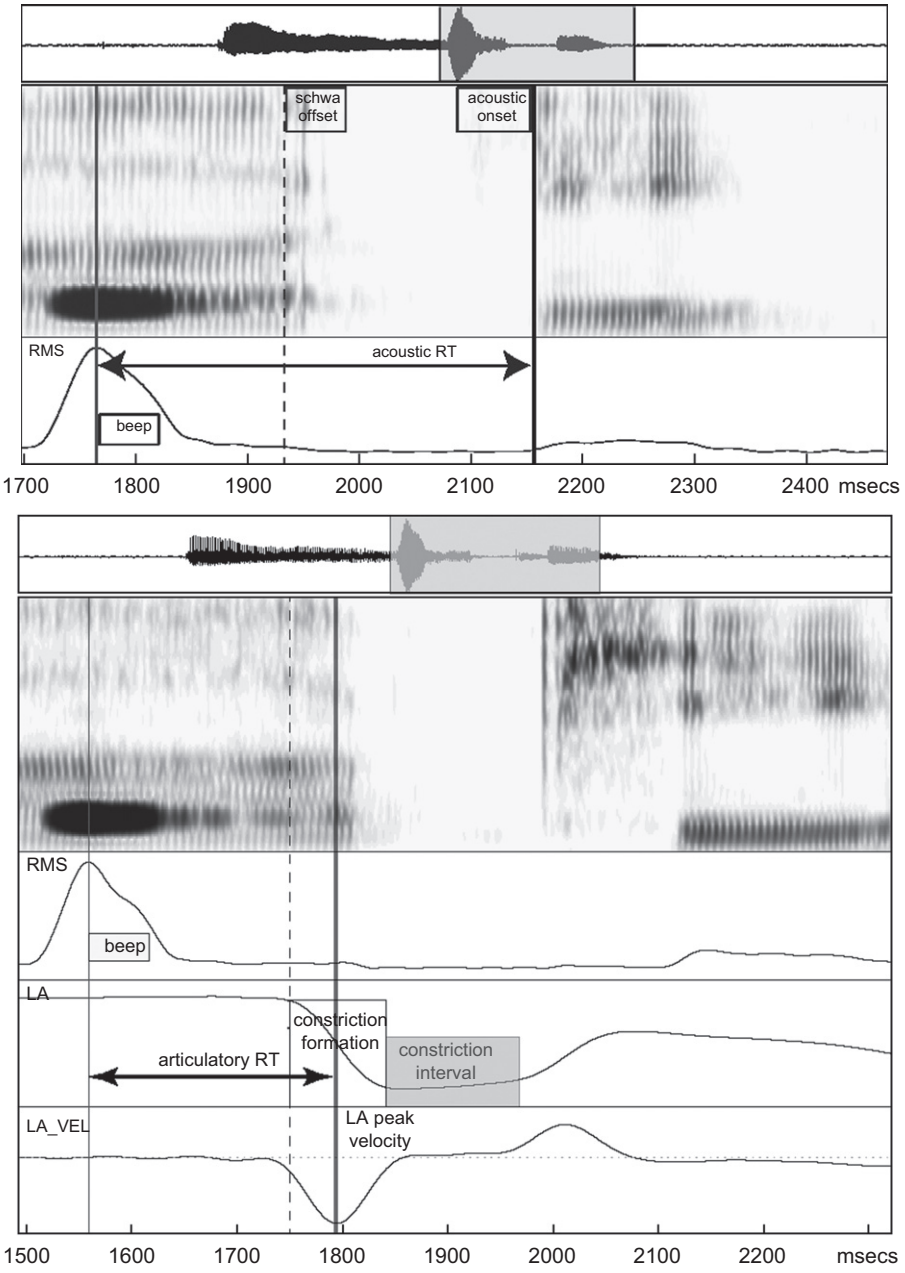

Fig. 2. Labeling of acoustic (upper figure) and articulatory data (lower figure). The upper figure exemplifies the acoustic labeling procedure for the item *ape* (acoustic signal, spectrogram, RMS for detecting the beep). The dashed line indicates the offset of the schwa and the bold line the acoustic onset for the vowel. The lower figure demonstrates the articulatory labeling criteria for the item *pea* (acoustic signal, spectrogram, RMS, lip aperture labeled 'LA' and velocity of LA, labeled 'LA_VEL'). For more details see text.

movements of the T1 signal. Due to the curved trajectory for dorsal stops (see e.g. Mooshammer, Hoole, & Kühnert, 1995) the vertical movement of T3 was used in 77% of all /k/s for measuring the gesture onset and 2D T3 for the remaining cases. For VC words with the diphthong /ei/ the offset of the first gesture could in 24% of all cases be more consistently detected by using the horizontal movement signal of T3. All other vowel gestures were labeled by using the 2D movements of the T3 signal. The peak of the beep was identified as the reference point for the onset. The interval between the peak of the beep and the articulatory peak velocity was used as the articulatory reaction time, henceforth RT_{art}. The peak velocity was chosen instead of gesture onset because it was temporally less variable than the gesture onset. The constriction formation duration of the initial gesture was calculated as the interval from the onset of the first constriction gesture to completion of that gesture measured as when the velocity fell below the 20% threshold (see white box in Fig. 2 for a bilabial constriction gesture). The constriction interval, i.e. the duration during which the articulator maintains its goal, is determined as the interval during which the velocity signal is below the threshold of 20% of the velocity peak (see grey box in Fig. 2).

Trials were excluded as early starts if gestural onset occurred less than 20 ms after the peak of the beep, which was the case for 4.67% of the total of 642 items. This exclusion criterion is shorter than for Experiment 1 because the articulatory onset RT_{art} preceded the acoustic onset RT_{ac} by a time that varied for different initial segments. Two items were produced with RT_{art} longer than 750 ms and were therefore excluded from the data set.

3.1.5. Statistics

As in Experiment 1, Linear Mixed Effects Models were calculated in order to test effects of syllable structure, consonant and vowel identity. As dependent variables we used RT_{art} and RT_{ac}. In order to get more stable results, the Markov Chain Monte Carlo sampling was performed with 10 000 simulations (instead of 1000 used for Experiment 1).

3.2. Results

Fig. 3 shows the means and standard error of RT_{ac} and RT_{art} for the four speakers of Experiment 2. Both measures show shorter reaction times for CV(C) than for VC for all consonant types. In the following section, RT_{ac} will be compared to the results of the previous experiment and then in the next section results for RT_{art} will be discussed.

3.2.1. Acoustic reaction times: RT_{ac}

Table 7 presents the Linear Mixed Effects Model analysis pooled across consonants (above) and split by consonants (below). Overall, RT_{ac} was significantly affected by STRUCT and CONS and the interaction between the two factors. RT_{ac} was affected significantly by STRUCT with reaction times for VC exceeding CV(C) by 39 ms ($\overline{x}_{VC} = 339$ ms, $\overline{x}_{CV(C)} = 300$ ms). This difference of RT_{ac} in the EMMA experiment was somewhat larger than that obtained in Experiment 1 (diff. = 29 ms). This divergence could either be due to some random factor like the selection of speakers or to a change in behavior due to the different experimental set-up. Since the participants in Experiment 2 also participated in Experiment 1, their results can be directly compared. First, RT_{ac} for the EMMA experiment was on average 36 ms longer. There was also an interaction between syllable structure and experiment: RT_{ac} of items with CV structure increased only by 22 ms whereas items with VC structure showed a 38 ms longer RT_{ac}. This result can be attributed to a single participant, who appeared to change her strategy between the two experiments. In Experiment 1, she produced very short or no pauses between the end of the preceding schwa and the vowel onset in VC sequences. She usually produced the transitions in a continuous manner with some glottalization. The duration from the offset of the schwa to the acoustic onset of the vowel in Experiment 1 was 74 ms for this speaker as compared to the other speakers who produced pauses and glottal stops in this context (175 ms, 135 ms, 157 ms). In Experiment 2, she more than doubled this duration to 162 ms by inserting pauses. Therefore, the larger differences between syllable types in the Experiment 2 than in Experiment 1 can be mainly attributed to this speaker's idiosyncratic change in strategy. There was also a marginal effect of CONS. As in Experiment 1, /s/ showed the shortest RTs. However, unlike Experiment 1, /l/ did not show an intermediate value between the stops and /s/.

Table 7
Results from Experiment 2: Linear Mixed Effects Model for RT_{ac}.

Subset	Effect	β	SE(β)	t-value	p_{MCMC}	Sig. diff.
All	Intercept	304.86	17.94			
	STRUCT	−20.10	3.89	−5.17	***	VC > CV(C)
	CONS	10.41	4.20	−2.48	*	/l/, stops > /s/
	STRUCT × CONS	13.38	5.18	2.58	*	
Stops	Intercept	313.46	18.64			
	STRUCT	−12.91	4.32	−2.99	***	VC > CV(C)
Lateral	Intercept	321.82	31.64			
	STRUCT	−54.28	12.08	−4.49	***	VC > CV(C)
Fricative	Intercept	252.42	16.68			
	STRUCT	−30.31	9.07	−3.34	**	VC > CV(C)

'.': $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

Apart from these subject specific differences there is also a more general difference that might explain why subjects tended to slow down during the EMMA experiments but not during the acoustic experiments. First of all, more repetitions were necessary for the EMMA experiment, and second, an intervening task was included in the EMMA experiment in order to avoid too much boredom. Therefore, the EMMA experiment took more time than the acoustic experiment. In general, slowing might lead to longer pauses in the case of VC stimuli, and thus the slowing could be expected to affect them differentially. In summary, results from the acoustic only experiment could not be fully replicated because the difference between CV and VC within in the stop subset was much smaller and did not reach significance in Experiment 1, but the same comparison was significant in Experiment 2. We assume that the increased number of repetitions, the slightly different material and a change in strategy for one speaker together might have yielded this discrepancy.

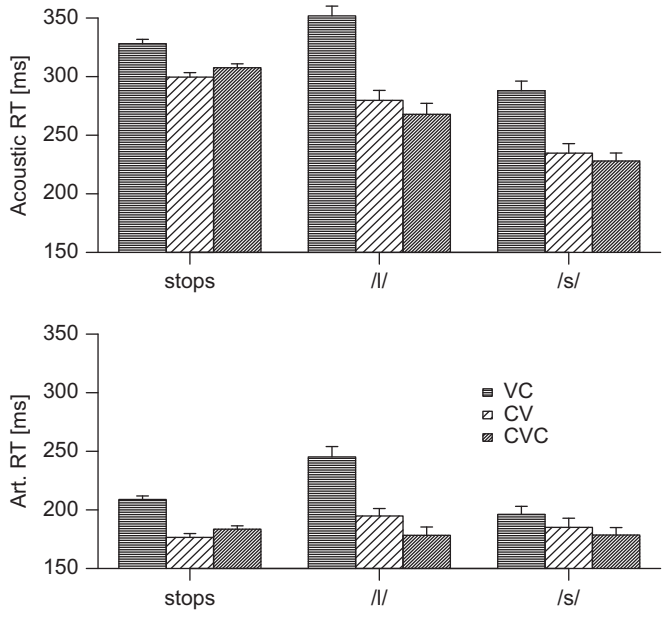

Fig. 3. Means and standard error of acoustically measured RT_{ac} (top) and the articulatorily measured RT_{art} (bottom) for different syllable structures, grouped by consonants. For each consonant type the subset consisted of VC syllables with identical coda consonants and of CV(C) syllables with identical initial onset consonants.

3.2.2. Articulatory reaction times: RT_{art}

For LME models for RT_{art} (measured as the interval from the GO signal to the velocity peak of articulatory movement for the initial gesture) 18 of 612 RT_{art} values were excluded as outliers. RT_{art} was significantly affected by the factor STRUCT (see Table 8, upper part), RT_{art} for VC being significantly longer than CV and CVC ($\overline{x}_{VC} = 215$ ms, $\overline{x}_{CV} = 180$ ms, $\overline{x}_{CVC} = 182$ ms). The factor CONS affected the articulatory reaction time significantly with longer RTs for items with /l/ and stops compared to /s/ ($\overline{x}_{/l/} = 212$ ms, $\overline{x}_{stops} = 191$ ms, $\overline{x}_{/s/} = 187$ ms). The post hoc tests, however, did not reach significance. The interaction between STRUCT and CONS was not significant, but in order to compare the results for RT_{ac} and RT_{art}, the data set was split by these two factors. The stops and the lateral showed a significantly longer RT for VC than for CV or CVC (stops: $\overline{x}_{VC} = 209$ ms, $\overline{x}_{CV} = 176$ ms, $\overline{x}_{CVC} = 183$ ms, lateral: $\overline{x}_{VC} = 245$ ms, $\overline{x}_{CV} = 195$ ms, $\overline{x}_{CVC} = 178$ ms) but no difference between CV and CVC. For items with /s/ the factor STRUCT was not significant ($\overline{x}_{VC} = 196$ ms, $\overline{x}_{CV} = 185$ ms, $\overline{x}_{CVC} = 179$ ms), but

Table 8
Linear Mixed Effects Model for RT_{art} for all sequences (upper part), and split into CV(C) sequences (middle part) and VC sequences (lower part).

Subset	Effect	β	SE(β)	t-value	p_{MCMC}	Sig. diff.
All	Intercept	194.40	12.63			
	STRUCT	−17.12	3.37	−5.23	***	VC > CV(C)
	CONS	−9.08	3.54	−2.56	*	
	STRUCT × CONS	6.93	4.36	1.59		
Stops	Intercept	192.22	12.75			
	STRUCT	−15.30	3.89	−3.94	***	VC > CV(C)
Lateral	Intercept	206.78	13.94			
	STRUCT	−34.64	8.76	−3.96	***	VC > CV(C)
Fricative	Intercept	187.66	15.27			
	STRUCT	−8.27	8.18	−1.01		
CV(C)	Intercept	181.63	11.39			
	CONS	3.90	4.43	0.88		
VC	Intercept	213.18	14.88			
	CONS	−25.73	8.14	−3.16	**	/l/ > /s/, stops

'.': $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

the differences between vowel-initial and consonant-initial items were in the expected direction. Adding a coda did not affect RT_{art} for any of the consonant sets.

Of further interest here was whether the initiation time differed for varying initial consonants within CV and CVC sequences as was found in Experiment 1 for RT_{ac}. As expected, results from Linear Mixed Effects Models, presented in Table 8, indicate that for RT_{art} the difference between initial consonants was much smaller than for Experiment 1 and did not reach significance ($\overline{x}_{/s/} = 182$ ms, $\overline{x}_{/l/} = 186$ ms, $\overline{x}_{stops} = 179$ ms). More details and possible explanations for this difference between RT_{art} and RT_{ac} will be discussed in the next section. RT_{art} of items with VC structure were significantly affected by the identity of the coda consonant with longer reaction times for the word *ale* than for VC words with stops or /s/ in the coda ($\overline{x}_{/s/} = 196$ ms, $\overline{x}_{/l/} = 245$ ms, $\overline{x}_{stops} = 209$ ms).

3.3. Discussion

Experiment 2 replicates the difference in initiation time between CV(C) and VC syllables, using direct articulatory measurement. Again, we consider the role of lexical and measurement factors in accounting for these results.

Role of lexical variables: As in Experiment 1 regressions between the RT_{art} values, averaged over repetitions and speakers, and lexical measures were calculated (see Table 9). Due to the low number of items and the fact that CONS was not significant for RT_{art}, data were not split by consonant as in Table 5. The results are quite similar to those for the acoustic data of Experiment 1. Only phone probability and biphone probability account for a substantial proportion of the variance in RT. The relationship is weaker and only marginally significant for biphone probability, unlike Experiment 1, where it was the strongest. This difference could be due to the different ranges of these variables across the materials in the two experiments, especially the lack of clusters (with very high biphone probabilities) in Experiment 2.

Again the relative importance of structural vs. lexical variables was tested via LME model selection. CV and CVC items were pooled together, such that the variable STRUCT had the two levels VC CV(C). Adding phone and biphone probability did not improve the model significantly ($\chi^2 = 2.44$, $p > 0.05$) compared to the model with only STRUCT. However, adding STRUCT did improve the model significantly compared to the model with only phone and biphone probability ($\chi^2 = 6.78$, $p < 0.01$). From this comparison it can be inferred that phone and biphone probability are less important than STRUCT for explaining the variation in RT_{art}.

Table 9
Slopes and adjusted R^2 for articulatory measured reaction time RT_{art} and several word-based measures. Data are averaged across speakers.

Characteristics	Slope	R^2
Word frequency	−2.5	−0.01
Neighborhood density	0.6	−0.03
Onset density	−0.5	−0.05
Phone probability	−719.23**	0.31
Biphone probability	−11,793	0.11
Syllable frequency	−2.3	0.01

'.': $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

If one of the probability variables were nonetheless driving the results, then we would expect a consistent effect of this variable within STRUCT categories. To test this, we chose phone probability (which was the only one with a significant regression with RT_{art} overall) and tested to see if that variable interacted with STRUCT in a linear model. A linear regression analysis was calculated for the averaged RT_{art} values with phone probability as a numerical predictor and STRUCT as a factorial predictor. The interaction between phone probability and STRUCT was indeed significant ($F(1, 17) = 18.71$, $p < 0.001$), suggesting that the effect of phone probability was not consistent across STRUCT categories. (In terms of overall main effects, phone probability was a significant predictor ($F(1,17) = 19.52$, $p < 0.001$) of RT_{art} whereas STRUCT by itself was not ($F(1, 17) = 1.3$, $p = 0.27$).) Fig. 4 helps visualize this interaction. The averaged RT_{art} values are plotted against the phone probability values. The bold line corresponds to the regression line for stimuli of all different syllables types. The nearly significant regression line ($F(1, 11) = 4.73$, $p = 0.053$) for CV and CVC pooled (light grey) are oriented in the same direction as the overall line, indicating that words with higher phone probability are initiated more quickly, as predicted by e.g. Vitevitch et al. (2004). For VC, however, the significant regression line ($F(1, 6) = 22.18$, $p < 0.01$) is oriented perpendicular to the other regression lines, indicating the opposite relationship.[1]

Even though it is not clear how to interpret this result, it underscores the importance of the purely structural variable. It also makes clear how important it is to include several syllable structures in experiments on speech planning. Most psycholinguistic studies focus on CVC words and almost all models of speech production are based on these results. Evidence for the necessity to include a wider variety of syllable structures within words in several kinds of lexicon-based word measures has been also presented by e.g. Kapatsinski (2005) and Brendel et al. (2011).

Role of measurement technique: In this section, the relationship between RT_{art} and RT_{ac} is compared across the set of initial phonetic segments, in order to better understand how such differences across segments can result from differences in the acoustic consequences of their initial gestures. The difference between RT_{art} and RT_{ac} is plotted in the white bars of Fig. 5, which show the means and standard errors of articulatory onset (measured at peak velocity) to acoustic onset interval (AAI) for different initial segments. Since adding a coda did not affect RT_{art} and RT_{ac}, only CV and VC segments will be compared. Voiceless consonants are plotted on the left, voiced segments (/l/ and vowels) on the right. As can be seen, there are substantial

[1] As was mentioned before, the VC item *ale* induced very long RT values. To test whether the exceptional behavior for this item could account for the significant interaction between phone probability and STRUCT a separate regression analysis excluding this item was calculated. Even though F values were smaller the interaction still reached significance ($F(1, 5) = 5.2$, $p < 0.05$).

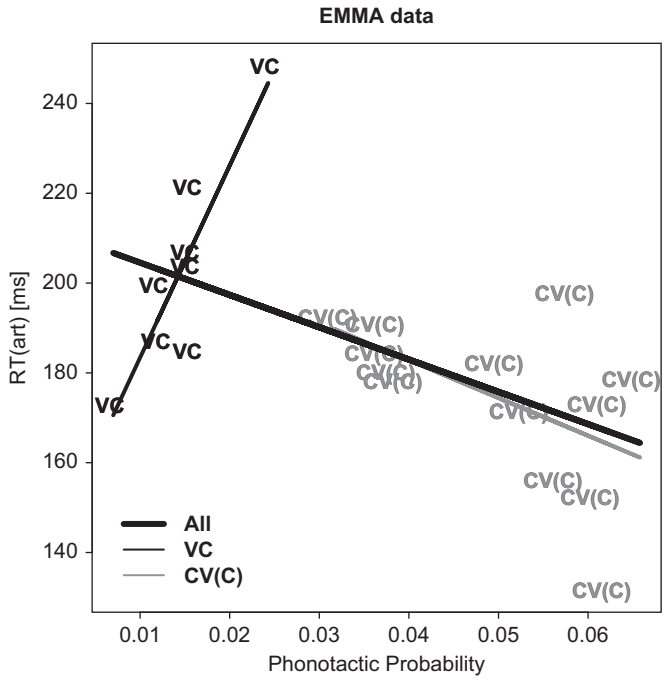

Fig. 4. Regression lines between phone probability and mean RT_{art} for Experiment 2.

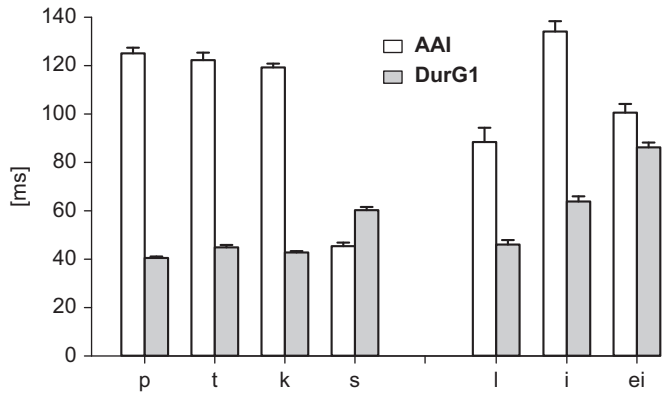

Fig. 5. Means and standard error of AAI (articulatory onset at peak velocity to acoustic onset interval, white bars) and DurG1 (constriction formation duration, grey bars) for different initial phonemes.

differences across initial segments from 134 ms for the initial vowel /i/ to 45 ms for /s/. One possible explanation of such differences would be that acoustic onset depends on the duration of the first gesture and that gestural duration varies across the segments. To help visualize this, the mean duration of constriction formation (DurG1, see white box in Fig. 2), starting from peak velocity, is plotted in the grey bars in Fig. 5. If this hypothesis is correct, then there should be a positive correlation between constriction formation duration and AAI, and the grey and white bars should show the same pattern of variation across segments. As can be seen, however, this not true. It is possible though to account for the differences in AAI and in the relation of AAI to DurG1 if we examine the voiceless and voiced segments separately.

For the voiceless consonants, the onset of acoustic energy should be mechanically (aero-acoustically) caused by the achievement of particular events in the normal articulation of these consonants in a /əː/ -VC context. In the case of stops, production of voiceless closure causes silence and there will be no energy in the signal until the oral constriction is released, which should take place considerably later than the achievement of its closure. This predicts that there should be a substantial difference between AAI and DurG1 for stops, which is supported in Fig. 5. All three stops show very similar patterns: similar values of AAI, and similar (much shorter) values of DurG1. For the fricative /s/, however, silence will not result during its production in this context, and the onset of turbulent energy (which was the criterion for acoustic onset) will occur when the constriction gets narrow enough. This has been shown in other work (Iskarous, Shadle, & Proctor, 2011; Mooshammer et al., 2006) to occur before the constriction gesture gets to its maximally constricted state, and we see that here in fact that for /s/ AAI is shorter than DurG1. Also, the fact that DurG1 is longer for /s/ than for stops is consistent with other data (Fuchs, Perrier, Geng, & Mooshammer, 2006).

For the voiced segments, in contrast, the acoustic onset is not mechanically caused by an event in their normal articulation in this context. The articulation of these segments following schwa does not necessarily produce any silence at all. However, in producing a sequence of vowel gestures with no intervening consonant, speakers of English will often produce a glottal constriction, resulting in either silence or glottalization (Pierrehumbert & Talkin, 1992), and reducing or eliminating the diphthongal percept that would otherwise result. In the case of the vowels in this experiment, subjects often chose to do this, and in the case of vowels, the release of this glottal constriction and onset of stable formant structure was taken as the onset of the vowel. However, as there is no causal mechanical relation between the achievement of the vowel gesture and the release of this glottal constriction, the glottal release would be expected to occur at some potentially variable time after the completion of the vowel gesture, and indeed we see that variability: for /i/, the glottal release occurs much later with respect to the achievement of the vowel gesture target than for /ei/. For /l/, the situation is more complex, as it was sometimes produced with glottalization or glottal closure (in which case that event would affect measurement of its acoustic onset), and other times it was not (in which case, the achievement of /l/ formant structure was used to measure its acoustic onset, and this is determined by the achievement of the /l/ gesture's target). As a result, we would expect that the value of AAI would be very variable for /l/, and as can be seen in the figure, the SEM is larger for /l/ than for any other segment (SEM /l/ 7.1 ms compared to next highest /s/ 5.3 ms, /t/ 6.0 ms).

In general, the data show that the technique of having talkers produce schwa during the delay interval successfully prevented them from beginning the articulation of the initial segment during the delay interval, even though the delay interval was long. Articulatory initiation was always observed to occur after the signal to respond. Without this constraint on articulation Kawamoto et al. (2008) have shown that early execution can occur, and one consequence of this early (silent) execution is that the difference in RT_{ac} between stops and fricatives disappears. Consistent with Kawamoto et al. (2008)'s analysis, the difference in RT_{ac} between stops and fricatives is observed here, where we can be sure early execution did not occur.

These results show that acoustically-measured onsets have a variable relation to the articulatory onset (which should be of primary relevance in studies of speech planning). While this relation is systematic in the case of voiceless consonants, it is more variable in the case of vowels and /l/. To evaluate the possibility that another acoustic landmark would be more systematically related to the articulatory events, we examined the end of the preceding schwa which was labeled in a wide-band spectrogram as visible changes in formants. In Fig. 6 the overall

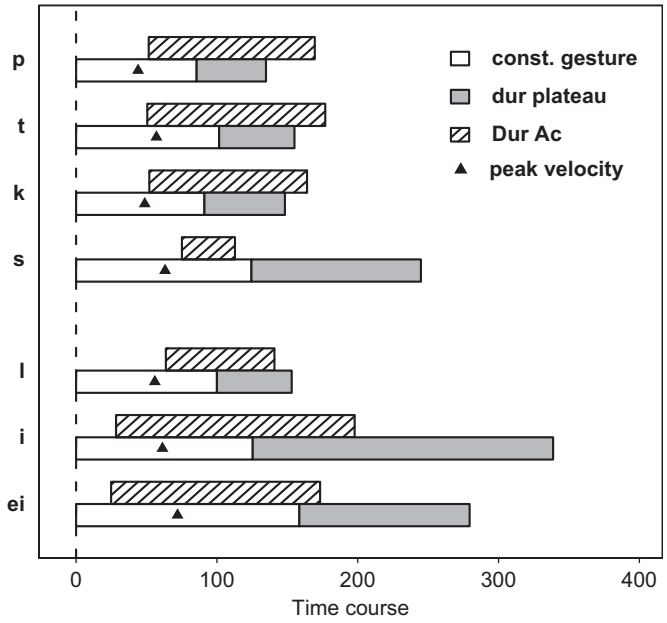

Fig. 6. Acoustic and articulatory landmarks relative to the articulatory onset for different initial phonemes. DurAc, shown as the striped bar, denotes the interval from a visible change in formant structure during the preceding schwa to the acoustic onset of the test stimulus. The white bar, termed const. gesture, represents the duration from the articulatory onset of the initial constriction gesture to the achievement of its target. The filled triangles correspond to the peak velocity during the first gesture. The grey bar shows the constricted interval, i.e. the interval from target achievement to target release.

means of the two acoustic landmarks and all the articulatory landmarks are shown lined-up with the onset of the first gesture (measured at the 20% threshold criterion, see Fig. 2). Again the graph is organized by voicing of the initial segment. For each initial segment, the upper (striped) bar shows the interval between the two acoustic landmarks, end of schwa and acoustic onset. The articulatory landmarks plotted are the onset of the initial gesture, the peak velocity of the gesture (black triangles; i.e. the event used to compute RT_{art}), the time it achieves its effective constriction target, and the time of constriction release. The interval from the gestural onset to the achievement of the constriction (=*constriction formation* in white box in Fig. 2) is shown as the white bar, the constricted interval (grey box in Fig. 2) is shown in the grey bar, both below the bar for the acoustic interval.

One regularity exhibited by the schwa offset landmark is that for all consonants (voiceless and /l/), the schwa offset occurs close in time (within a range of −12 ms for /s/ and 6 ms for /t/) to the time of peak velocity of the initial gesture. Since movements associated with the gesture will cause changes in vocal tract resonances, and since the schwa end for consonants was defined in terms of consonant-directed changes in formants, this is reasonable. At the time of peak velocity, the rapid changes in vocal tract shape should be sufficient to have a noticeable effect on the formants. However, this was not the case for the vowels: here the schwa offset precedes the velocity peak substantially (with 47 ms for the diphthong and 33 ms for /i/). The reason for this is that for vowels, the end of voicing caused by glottal constriction sometimes preceded the onset of formant transitions, so that no formant transitions were visible. In these cases, the schwa end was defined by the end of voicing. This measure of schwa offset must necessarily be earlier than one based on the transitions, which reflect the time of peak velocity. Thus, the schwa offset landmark also does not reflect some fixed point in the articulatory trajectory.

4. General discussion

The results of Experiments 1 and 2 demonstrate a small, but robust, difference in the time required to initiate articulation for CV vs.VC syllables. This can be interpreted as supporting the hypothesis temporal coordination can be established more quickly or efficiently for CV organization of gestures than for VC organization of gestures. However, there are some alternative accounts of the results that do not make reference to these organizational properties at all, and we first consider these below. We then consider possible syllable structure accounts for the observed effects.

Glottalization accounts: One key property of the VC productions that might contribute to an alternative account of the results is that they were often produced with an initial constriction gesture of the glottis, which resulted in either a glottal stop or glottalization in the acoustic signal. This is a regular process in English, most reliably occurring in initial vowels following an Intonational Phrase Boundary (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996; Pierrehumbert & Talkin, 1992), which is the case for responses in this experiment (there is an IP boundary between the schwa and the target word). There are three different ways in which the presence of this glottal gesture might be expected to slow the measured response time for VC stimuli. First, the glottal gesture and its timing must be planned, so VC stimuli require one more gesture to be planned than the comparable CVs, and this could require more planning time. To address this question, the data of Experiment 2 were checked for the existence of glottalization. Because it is impossible to distinguish between multiple glottal stops and glottalization, vowel and lateral initial items were labeled as glottalized if one or more distinct glottal stops were present or if irregular phonation was visible in the spectrogram. Before outlier exclusion, 137 out of 236 VC words were produced with glottalization. As was expected from the literature (see e.g. Dilley et al., 1996) speakers vary a great deal in their propensity to use glottalization: two speakers glottalized almost 80% of all VC items whereas the other two speakers only glottalized about 34%. Assuming that glottalized items are produced with an additional glottal constriction gesture and this additional gesture is slowing down planning, then longer RTs are expected for these items. However, mean RT_{art}s of VC items do not differ whether they are glottalized (211 ms) or not (212 ms). Subsetting the data according to the coda consonant showed that laterals and fricatives had shorter RT_{art}s for glottalized items whereas stops had longer RTs. Therefore, additional gestures do not necessarily require more planning time before the production is initiated. Further, in order to test whether the longer response lags for VCs than for CV(C)s were obtained even in the absence of glottalization, a Linear Mixed Effects model was computed for the data excluding glottalized VC items. Results indicate that modal voice VCs are initiated 33 ms later than CVs ($t=-4.4$, $p<0.001$) and 30 ms later than CVCs ($t=-3.6$, $p<0.001$).

A second possible account would be to suggest that the appropriate time point at which to measure RT_{art} is the peak velocity of the glottal constriction (which could not be measured) and if initiation of the vowel constriction gesture (which was actually used to measure RT_{art}) occurs later than the initiation of the glottal gesture, this could account for the apparently slow RTs for VCs. However, this account would also not predict the results obtained for VCs lacking glottalization as just discussed.

Finally, a third account would be that the glottal constriction is more demanding of planning resources than other gestures, perhaps because it is a prosodically, rather than lexically, controlled gesture for English, and because it is variable in its occurrence. However, as noted in the Discussion of Experiment 2, the CV(C) stimuli with /l/ were sometimes also produced with

irregular glottal pulses (five out of 53 items). This would predict a somewhat smaller VC–CV(C) difference for /l/ words, than for stops or fricatives. Yet the VC–CV(C) difference in RT_{art} was actually larger than for stops or fricatives, so this account is also not likely to be correct.

A related account of the VC–CV difference could focus on the hiatus (vowel sequence) environment created between the schwa and the VC. Since hiatus in English is avoided (e.g. McCarthy, 1993), this could make the /əː/ -VC sequence difficult to produce. However, as noted above (see Section 2.1.3), the results of the delayed naming condition without preceding schwa did not differ from the results of the schwa condition, which makes this account seem unlikely.

P-center account: When speakers (or listeners) are asked to align a sequence of syllables at equal temporal intervals, it is not the acoustic onsets of the syllables that they appear to align. Rather, they align a point close to the acoustic vowel onset, which has been dubbed the perceptual center (or p-center) of the syllable (see Fowler, 1979; Marcus, 1981; Pompino-Marschall, 1989). Even though there are problems in identifying the speech event(s) that align most closely with the p-center across different conditions (for an extensive and recent review see Villing, Repp, Ward, & Timoney, 2011), it has been found that the acoustic vowel onset (see Janker, 1996), or better, the peak velocity of the vowel gesture (see de Jong, 1994) seems to be reasonable approximations. The results reported here bring to mind the p-center as a possible alternative explanation. If speakers are aligning the p-center of their produced syllables a fixed time from the GO signal, this would predict that the time from the GO signal to the measured (acoustic or articulatory) onset of the syllable would be shorter for CV(C) than for VC syllables, which is the main finding reported here, and also the time would be shorter for CCV(C) than for CV(C), which was also accords with the results. However, this account is unlikely for at least three reasons.

First and most importantly, the task in this experiment is explicitly *not* a regular rhythmic one of the sort that has elicited the p-center effect. The time from the stimulus presentation is randomized, and it would be impossible to impose a regular rhythm on the sequence of GO signals. There is no evidence that the p-center would be relevant to a task like this. Second, if subjects were aligning the p-center to the GO signal, then a point near the onset or peak velocity of the vowel gesture should be timed *relatively* invariantly with respect to the GO signal, compared to the difference in lag between CV and VC stimuli when measured from their respective (C or V) onsets. Because of blending between vowel and consonant gestures in most combinations, an articulatory analysis of the onset of the vocalic gesture in CV sequences could only be done for initial bilabial stops for which the consonantal and the vocal articulators show the least amount of overlap. To test this, we measured the articulatory peak velocity of the vowel gesture with respect to the GO signal for both /pV/ and /Vp/ stimuli in Experiment 2. The results show that the peak velocity point is 33 ms later in /pV/ than in /Vp/ stimuli, so this point in time is *not* being invariantly timed to the GO signal. In fact, 33 ms is *less* invariant than the 21 ms difference observed using peak velocities of the initial C gestures for /pV/ vs. the initial V gestures for /Vp/ stimuli (the actual RT_{art}). Third, it is also known that temporal properties of the syllable later than the vowel onset also affect the p-center (de Jong, 1994; Villing et al., 2011). Thus, we would expect small but systematic differences in alignment of the p-center in CV vs. CVC stimuli, which we did not observe. Overall, it does not appear that a p-center account of the results is very promising.

Syllable structure accounts: Having argued that several alternative accounts are probably inadequate to account for the results of the experiments reported here, we conclude that results can be interpreted as tentatively supporting the hypothesis that assembling a gestural plan for V(C) syllables requires more time than for CV(C) syllables. Here we consider why this might be the case.

One hypothesis is that speakers have more experience producing CV(C) syllables than V(C) syllables, and that gestural timing plans for commonly produced (high probability) sequences are well-learned and are either stored (Levelt & Wheeldon, 1994) or can be more quickly assembled. The fact that the phonotactic probability of the V(C) stimuli was significantly less than that of the CV(C) stimuli supports this analysis, as does the fact that a large proportion of the cross-item variability in RT (in both experiments) can be predicted from phonotactic probability (either phone or biphone or both). These probability facts are themselves consistent with the typological properties and markedness of these syllable types: the preference for syllables to have onsets (Clements, 1990; Clements & Keyser, 1983; Prince & Smolensky, 1993). However, the results suggest that the source of the effect is in the probabilities themselves, not in the universal markedness properties. Stimuli with initial clusters of Experiment 1 (CCV(C)) exhibited shorter RTs than CVs, which is consistent with the fact that the biphone probability of the clusters was significantly higher than that of the CVs, but is not consistent with the fact that CCV structures are more marked than CV.

One issue that can be raised about the probability effects on planning time is whether experience producing a particular sequence determines the planning time for that sequence, or whether due to some kind of generalization during learning, the probability of a syllable structure class is what controls planning time. This is similar to questions raised by Steriade (2001) in phonological theory about whether assimilation and neutralization phenomena that appear to be syllable-structure dependent are in fact governed not by syllable structure *per se*, but rather by the perceptual or production characteristics of particular segmental sequences (her analyses support the non-structural alternative). The results of the current study are ambivalent on this point. The stepwise regression analyses and the within-syllable structure category regressions performed on the output of Experiment 2 slightly favor the generalized structural effects, but those results are not particularly strong, and a new experiment with wider sampling of within-category probabilities would be needed to resolve the issue.

Exactly how it is that sequence probability and/or syllable structure influence initiation time is, of course, of great interest. It is appealing to implicate the planning of inter-gestural timing for the gestures of the first two segments, as the current results are compatible with, and partly predicted by, the coupled oscillator planning model (Goldstein et al., 2006; Nam, 2007; Nam & Saltzman, 2003; Saltzman et al., 2008), the most explicitly worked out model for planning the timing of the gestures of a syllable. The planning oscillators (clocks) that are responsible for triggering the gestures are set into motion at the outset of planning, with each oscillator initialized to some (possibly arbitrary) phase. Due the inter-oscillator coupling imposed by the syllable structure graph, the oscillators settle into stable attractors of relative phase. The effect of experience on the process could be modeled though the phase initialization. The planning system could learn the distribution of relative phases that resulted from previous instances in which a given syllable was planned and use the mean of that distribution to set the initial phase conditions for the current planning episode. The closer the initial phases are to the stable attractor values, the faster the settling process will be. By using experience to guide phase initialization, the oscillators for well-learned syllables will settle more quickly, and gestures can be initiated more quickly. This suggestion is much in the spirit of exemplar models of phonological representation

(Pierrehumbert, 2001, 2003), and particularly the dynamical field model of experience on assembling phonological representations (Gafos & Kirov, 2009), which makes explicit connections between experience and the time to assemble phonetic control structures.

The coupled oscillator planning model also makes the prediction that there should be an intrinsic difference in settling time as a function of syllable structure, even for syllables with an equivalent level of experience. The oscillators for a CV structure are predicted to settle more quickly than those for a VC structure, because the in-phase coupling of CVs is effectively stronger than the anti-phase coupling in VCs. Unfortunately, it is not possible to directly test this prediction in this experiment, because (as can be seen in Fig. 4), the ranges of phonotactic probability for VC and CV(C) stimuli are non-overlapping. However, it is unlikely that this intrinsic effect is responsible for the major differences that were found, because the RT for bare V stimuli in Experiment 1 were the same as for VC stimuli, and there is no anti-phase coupling required for the V stimuli. Nonetheless, the stronger coupling of in-phase might be a factor contributing to the preference for CV syllables in languages, which in turn causes the differences in planning times that we observed here. The perceptual salience of the consonant release in CVs has also been proposed as a basis for the preference for onsets vs. codas, both across languages and within languages (Ohala, 1996; Steriade, 2001). It is not clear whether this account could also explain why onset-less syllables of the form V (no onset or coda) are also disfavored. A possible account for this could be found in the excess degrees of articulator freedom found in V as opposed to CV syllables. In CV syllables, the C and V gestures are planned (and executed) concurrently. The simultaneous demands of C and V on the articulator system will constrain the articulators to relatively narrow ranges that can be used accomplish the joint task. In a V syllable, there is a wider range of articulatory possibilities, so the choice of articulator values has more uncertainty associated with it, which might slow down its planning.

5. Conclusions

The experiments reported here show that the time required to initiate articulation of CV structures is shorter than for VC structures. The experimental design and post-hoc analysis make it unlikely that the results are due to differences in speed of lexical access or the fact that VC structures can be articulated with initial glottalization. The results can be explained by differences in phonotactic probability of CV and VC sequences, possibly through the effect that these variables have on planning the gestural timing for the initial portion (two segments) of the sequence. This account is consistent with the coupled oscillator model of planning gestural timing. The experiments also illustrate the utility of combining reaction time methods with measurement of articulatory kinematics, a combination that will be necessary for a more complete understanding (and model) of speech production, particularly the regularities of gestural timing.

Acknowledgments

Research was supported by NIH NIDCD DC008780 to Haskins Laboratories. Thanks to Manisha Kulshreshtha and Man Gao for assistance with running the experiments and analyzing the data.

References

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge University Press.

Baayen, H., Piepenbrock, R., & Rijn, H. (1995). *The CELEX lexical database (release 2) [CD-ROM]. Linguistic Data Consortium*. Philadelphia: University of Pennsylvania.

Bell-Berti, F., & Harris, K. (1981). A temporal model of speech production. *Phonetica*, *38*, 9–20.

Bohland, J., Bullock, D., & Guenther, F. (2009). Neural representation and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience*, *22*, 1504–1529.

Brendel, B., Erb, M., Riecker, A., Grodd, W., Ackermann, H., & Ziegler, W. (2011). Do we have a "Mental Syllabary" in the brain? An fMRI study. *Motor Control*, *15*, 34–51.

Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology*, *3*, 219–252.

Browman, C., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, *45*, 140–155.

Browman, C., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*, 201–251.

Browman, C., & Goldstein, L. (1990a). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, *18*, 299–320.

Browman, C., & Goldstein, L. (1990b). Tiers in articulatory phonology, with some implications for casual speech. In: M. Beckman, & J. Kingston (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 341–376). Cambridge: University Press.

Browman, C., & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structure. *Les Cahiers de l'ICP, Bulletin ICP*, *5*, 25–34.

Byrd, D. (1995). C-centers revisited. *Phonetica*, *52*, 285–306.

Byrd, D., & Tan, C. (1996). Saying consonant clusters quickly. *Journal of Phonetics*, *24*, 263–282.

Byrd, D., Tobin, S., Bresch, E., & Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *Journal of Phonetics*, *37*, 97–110.

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

Clements, G. (1990). The role of the sonority cycle in core syllabification. In: J. Kingston, & E. M. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 283–333). Cambridge: Cambridge University Press.

Clements, G., & Keyser, S. J. (1983). *CV phonology*. Cambridge, MA: MIT Press.

Cohn, A. (1993). The status of nasalized continuants. In: M. Huffman, & R. Krakow (Eds.), *Nasal, nasalization, and the velum* (pp. 329–367). Academic Press.

Dell, S. G. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.

Dell, S. G., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*, 287–314.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*, 423–444.

Fowler, C. A. (1979). "Perceptual centers" in speech production and perception. *Perception & Psychophysics*, *25*, 375–388.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*, 113–133.

Fuchs, S., Perrier, P., Geng, C., & Mooshammer, C. (2006). What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents. In: J. Harrington, & M. Tabain (Eds.), *Towards a better understanding of speech production processes* (pp. 149–164). New York: Psychology Press.

Fujimura, O. (1981). Temporal organization of articulatory movements as a multidimensional phrasal structure. *Phonetica*, *38*, 66–83.

Fujimura, O. (1986). Relative invariance of articulatory movements: An iceberg model. In: S. J. Perkell, & H. D. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 226–242). Hillsdale: Erlbaum.

Gafos, I. A., & Kirov, C. (2009). A dynamical model of change in phonological representations: The case of lenition. In: F. Pellegrino, E. Marsico, I. Chitoran, & C. Coupé (Eds.), *Approaches to phonological complexity* (pp. 219–240). Mouton de Gruyter.

Gao, M. (2008). *Tonal alignment in Mandarin Chinese*. Ph.D. Thesis. Yale University.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldinger, S., Tamiko, A., Abramson, M., & Jain, P. (1997). Open wide and say "blah!": Attentional dynamics of delayed naming. *Journal of Memory and Language*, *37*, 190–216.

Goldrick, M., & Blumstein, S. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*, 649–683.

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In: M. A. Arbid (Ed.), *From action to language: The mirror neuron system* (pp. 215–249). Cambridge: Cambridge University Press.

Goldstein, L., Chitoran, I., & Selkirk, E. (2007a). Syllable structure as coupled oscillator modes: Evidence from Georgian vs. Tashlhiyt Berber. In: *Proceedings of the XVI international congress of phonetic sciences* (pp. 241–244).

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007b). Dynamic action units slip in speech production errors. *Cognition*, *103*, 386–412.

Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. In: G. Fant, H. Fujisaki,

& J. Shen (Eds.), *Festschrift for Wu Zongji* (pp. 239–250). Beijing: Commercial Press.
Grainger, J., Spinelli, E., & Ferrand, L. (2000). Effects of baseword frequency and orthographic neighborhood size in pseudohomophone naming. *Journal of Memory and Language*, *42*, 88–102.
Guenther, F. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594–621.
Haken, H., Kelso, J., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*, 347–356.
Harley, T. A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes*, *8*, 291–309.
Hermes, A., Grice, M., Mücke, D., & Niemann, H. (2008). Articulatory indicators of syllable affiliation in word initial consonant clusters in Italian. In: R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the eighth international seminar on speech production* (pp. 433–436).
Hermes, A., Ridouane, R., Mücke, D. & Grice, M. (2011). Kinematics of syllable structure in Tashlhiyt Berber: The case of vocalic and consonantal nuclei. In D. Ostry, S. Baum, L. Menard, & V. Gracco (Eds.), *Proceedings of the ninth international seminar on speech production* (pp. 401–408). Montreal, Canada.
Honorof, D., & Browman, C. P. (1995). The center of edge: How are consonant clusters organized with respect to the vowel?. In: K. Elenius, & P. Branderud (Eds.), *Proceedings of the XIIIth international congress of phonetic sciences* (pp. 552–555).
Iskarous, K., Shadle, C. H., & Proctor, M. (2011). Articulatory-acoustic kinematics: The production of American English /s/. *Journal of the Acoustical Society of America*, *129*, 944–954.
Janker, P. (1996). Evidence for the P-center syllable nucleus-onset correspondence hypothesis. *ZAS Papers in Linguistics*, *7*, 94–124.
Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 824–843.
de Jong, K. (1994). The correlation of p-center adjustments with articulatory and acoustic events. *Attention, Perception & Psychophysics*, *56*, 447–460. doi:10.3758/BF03206736.
Kapatsinski, V. (2005). Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. *Research on Speech Perception, Progress Report*, *27*, 134–152.
Kapatsinski, V. (2010). Frequency of use leads to automaticity of production: Evidence from conversation. *Language and Speech*, *53*, 71–105.
Kawamoto, A., & Kello, C. (1999). Effect of onset cluster complexity in speeded naming: A test of rule-based approaches. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 361–375.
Kawamoto, A., Lui, Q., Mura, K., & Sanchez, A. (2008). Articulatory preparation in the delayed naming task. *Journal of Memory and Language*, *58*, 347–365.
Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, *47*, 145–171.
Krakow, R. (1993). Nonsegmental influences on velum movement patterns: Syllables, sentences, stress, and speaking rate. *Phonetics and Phonology*, *5*, 87–116.
Kühnert, B., Hoole, P., & Mooshammer, C. (2006). Gestural overlap and C-center in selected French consonant clusters. In: H. C. Yehia, D. Demolin, & R. Laboissière (Eds.), *Proceedings of the seventh international seminar on speech production* (pp. 327–334).
Levelt, W. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition*, *42*, 1–22.
Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.
Levelt, W., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*, 239–269.
Löfqvist, A., & Gracco, V. (2002). Control of oral closure in lingual stop consonant production. *Journal of the Acoustical Society of America*, *111*, 2811–2827.
Lukatela, G., & Turvey, M. (2000). An evaluation of the two-cycles model of phonology assembly. *Journal of Memory and Language*, *42*, 183–207.
Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics*, *30*, 247–256.
Marin, S. (2011). Organization of complex onsets in Romanian. In D. Ostry, S. Baum, L. Menard, & V. Gracco (Eds.), *Proceedings of the ninth international seminar on speech production* (pp. 179–186). Montreal, Canada.
Marin, S., & Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control*, *14*, 380–407.
McCarthy, J. (1993). A case of surface constraint violation. *Canadian Journal of Linguistics*, *38*, 169–195.
Monsell, S. (1990). Frequency effects in lexical tasks: Reply to Balota and Chumbley. *Journal of Experimental Psychology: General*, *119*, 335–339.
Mooshammer, C., & Fuchs, S. (2002). Stress distinction in German: Simulating kinematic parameters of tongue-tip gestures. *Journal of Phonetics*, *30*, 337–355.
Mooshammer, C., Hoole, P., & Geumann, A. (2006). Interarticulator cohesion within coronal consonant production. *Journal of the Acoustical Society of America*, *120*, 1028–1039.
Mooshammer, C., Hoole, P., & Kühnert, B. (1995). On loops. *Journal of Phonetics*, *23*, 3–21.
Mulatti, C., Reynolds, M. G., & Besner, D. (2006). Neighborhood effects in reading aloud: New findings and new challenges for computational models. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 799–810.
Nam, H. (2007). Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. In: J. Cole, & I. J. Hualde (Eds.), *Laboratory phonology*, Vol. 9 (pp. 483–506). New York: Mouton de Gruyter.
Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. In: M.-J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th international congress of phonetic sciences* (pp. 2253–2256). Barcelona: Futurgrafic.
Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1718–1725.
Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*, 273–281.
Perkell, S. J. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. York, PA: Maple Press.
Perkell, S. J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992). Electro-magnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America*, *92*, 3078–3096.
Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In: J. Bybee, & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137–157). John Benjamins.
Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*, 115–154.
Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In: G. Docherty, & R. D. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 90–127). Cambridge: University Press.
Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). *Synchronization: A universal concept in nonlinear sciences*. Cambridge University Press.
Pinheiro, J., & Bates, M. D. (2000). *Mixed effects models in S and S-PLUS*. New York: Springer-Verlag.
Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics*, *17*, 175–192.
Prince, A., Smolensky, P. (1993). Optimality theory: constraint interaction in generative grammar. TR 2 Rutgers Center for Cognitive Science.
Rastle, K., Croot, K., Harrington, J., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 1083–1095.
Saltzman, E., & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, *19*, 499–526.
Saltzman, E., & Munhall, G. K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*, 333–382.
Saltzman, E., Nam, H., Krivokapić, J., & Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In: P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the speech prosody 2008* (pp. 175–184).
Shaw, J., Gafos, I. A., Hoole, P., & Zeroual, C. (2009). Syllabification in Moroccan Arabic: Evidence from temporal stability in articulation. *Phonology*, *26*, 187–215.
Steriade, D. (2001). Directional asymmetries of place assimilation: A perceptual account. In: E. Hume, & K. Johnson (Eds.), *The role of speech perception in phonology* (pp. 219–250). San Diego, CA: Academic Press.
Sternberg, S., Knoll, R., Monsell, S., & Wright, C. (1988). Motor programs and hierarchical organization in the control of rapid speech. *Phonetica*, *45*, 175–197.
Sternberg, S., Monsell, S., Knoll, R., & Wright, C. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In: G. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 117–152). New York: Academic Press.
Tabain, M., Breen, G., & Butcher, A. (2004). VC vs. CV syllables: A comparison of aboriginal languages with English. *Journal of the International Phonetic Association*, *34*, 175–200.
Turvey, M. (1990). Coordination. *American Psychologist*, *45*, 938–953.
Villing, R., Repp, B., Ward, T., & Timoney, J. (2011). Measuring perceptual centers using the phase correction response. *Attention, Perception & Psychophysics*, *73*, 1614–1629.
Vitevitch, M. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 735–747.
Vitevitch, M., Armbruester, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 514–529.
Vitevitch, M., & Luce, P. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, *36*, 481–487.
Vousden, J. I., Brown, G. D. A., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology*, *41*, 101–175.
Zsiga, E. (1995). An acoustic and electropalatographic study of lexical and post-lexical palatalisation in American English. In: B. Connell, & A. Arvaniti (Eds.), *Phonology and phonetic evidence: Papers in laboratory phonology IV* (pp. 282–302). Cambridge University Press.