A Unified Account of Segment Duration and Coarticulatory Effects in Speech Production

Alan H. Kawamoto (ahk@ucsc.edu)
Department of Psychology, University of California at Santa Cruz
Santa Cruz, CA 95064 USA

Qiang Liu (qxliu@ucsc.edu)
Department of Psychology, University of California at Santa Cruz
Santa Cruz, CA 95064 USA

Abstract

In this paper, we describe a unified account of anticipatory coarticulatory effects and local segment duration effects based on a segmental minimal unit of articulation. The anticipatory coarticulatory effects are modeled using a Jordan network and the segment duration effects are modeled using Sigma-Pi control units. The viability of this approach in resolving an ongoing debate between a segmental versus syllabic minimal unit of articulation in speech production is illustrated using a small lexicon.

Keywords: Coarticulation; Phonological Priming; Word Naming; Speech Production; Segment Duration; Jordan Network; Minimal Unit of Articulation

Minimal Unit of Articulation

Generally, the final stage of speech production is believed to involve two broad processes: (1) phonological encoding, in which an abstract representation (i.e., phonological code) of speech is generated, and (2) articulation, in which the phonological code is realized as overt speech. Despite wide acceptance of this framework, debate is still ongoing with respect to the nature of the minimal unit involved in articulation. Some researchers favor the phonological word (e.g., Levelt, Roelofs, & Meyer, 1999) or the syllable (e.g., Meyer, Roelofs, Levelt, 2003), while others favor the segment (e.g., Kawamoto, Kello, Jones, & Bame, 1998; Mackay, 1987).

Arguments for the syllable have typically been based on the premise that anticipatory coarticulation, the influences of a subsequent segment on the production of the current segment, is a necessary and obligatory accompaniment of speech (Fujimura & Lovins, 1978). Thus, articulation cannot be initiated on the basis of anything smaller than the syllable because the phonetic environment of a segment must be known for anticipatory coarticulation (Levelt et al., 1999; Rastle, Coltheart, Harrington, & Palethorpe, 2000).

By contrast, Kawamoto and his colleagues have assumed that the segment is the minimal unit of articulation (Kawamoto, 1999; Kawamoto, Kello, Jones, & Bame, 1998). Arguments for the segment have generally been based on local segment duration effects, particularly segment duration increases that are a result of local processing difficulties in generating the phonological code of a subsequent segment (see Kawamoto, 1999; Kawamoto et al., 1998).

Previous Attempt to Reconcile Differences

One way to reconcile the differences is to assume that the minimal unit of articulation is the segment and that different criteria to initiate articulation are used (Kawamoto et al., 1998; Kello, 2004). That is, an individual could choose to initiate articulation as soon as the initial segment is encoded (i.e., the segment criterion) or only after the full phonological code for the whole word becomes available (i.e., the whole word criterion). When the whole word criterion is used, anticipatory coarticulation would be produced to its full extent without any local segment duration effects. The same would be true, regardless of the criterion used, if the phonological code for the syllable becomes available all at once. By contrast, when articulation is initiated using the segment criterion without the complete phonological code, anticipatory coarticulation would only be produced when the relevant segment (e.g., the vowel) becomes available. Moreover, the articulation of the segment immediately preceding the locus of processing difficulty would be lengthened until the processing difficulty is resolved.

On the other hand, if the minimal unit of articulation is the syllable, then the notion of an articulation criterion becomes moot for monosyllabic words because the entire phonological code would be available when articulation is initiated. Under this assumption, there would be no opportunity for local segment duration effects to be manifested, and anticipatory coarticulation would always be produced to its full extent. Note that these results are indistinguishable from a segmental minimal unit of articulation and a whole word criterion.

Articulatory Differences in Different Priming Conditions

It turns out that coarticulation is not as simple as depicted above by the syllabic perspective. The same response can be produced with or without anticipatory coarticulation under different phonological prime conditions (Kawamoto, Liu, Kherlein, & Johnson, 2007; Whalen, 1990). In Whalen's study, participants produced four VCV utterances, all beginning with the same vowel, in two experimental conditions. In one, participants knew beforehand what the consonant was but not the second vowel, and in the other, they knew what the second vowel

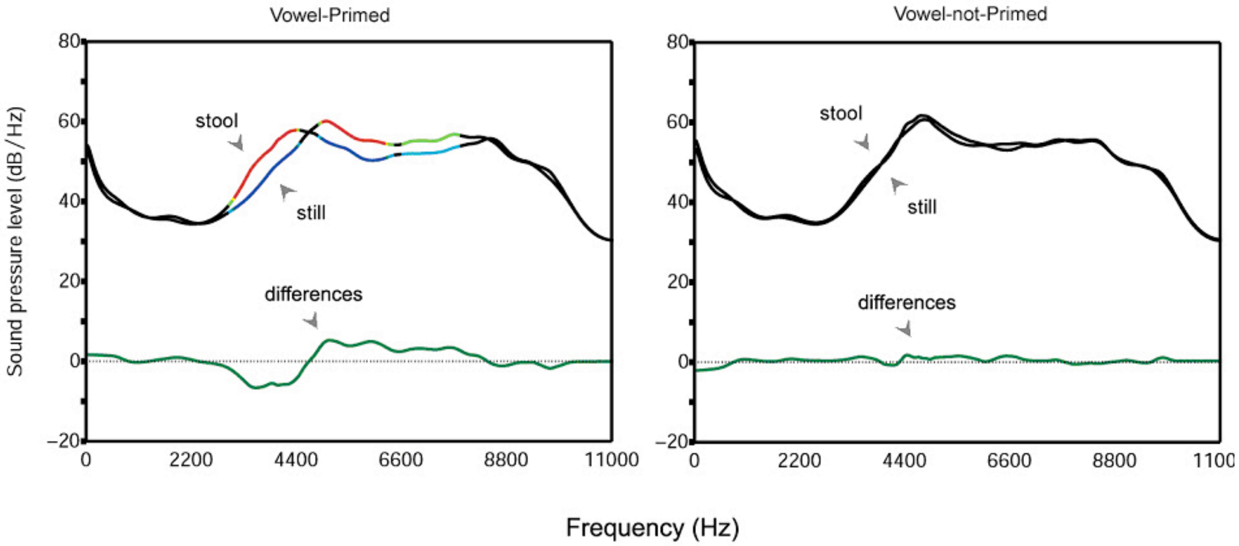

Figure 1: The mean fricative spectrum measured at the midpoint of the /s/ for the tokens of *stool* and *still*, for one participant, produced in the Vowel-Primed condition (left) and the Vowel-not-Primed condition (right).

was but not the consonant. Whalen found well-defined anticipatory vowel-to-vowel coarticulation effects between syllables only when the vowel was primed.

Kawamoto et al. (2007) extended Whalen's results to anticipatory vowel-to-consonant coarticulation within a syllable. They had participants produce four monosyllabic CCVC words, *spill*, *spool*, *still*, and *stool*, under one of two priming conditions, Vowel-Primed and Vowel-not-Primed. In the Vowel-Primed condition, participants produced tokens of one of two 2nd segment minimal pairs (i.e., *spill-still* or *spool-stool*), and in the Vowel-not-Primed condition, participants produced tokens of one of two vowel minimal pairs (i.e., *spill-spool* or *still-stool*). These conditions created two situations where the segments of a word, except for the divergent segments, were phonologically primed.

To allow participants their choice of criterion to initiate articulation, the standard naming task was used (see Rastle et al., 2000). Kawamoto and colleagues found that the majority of participants exhibited extensive anticipatory coarticulation differences between the tokens of a particular vowel minimal pair in both conditions. However, a small number of participants only showed extensive anticipatory coarticulation differences in the Vowel-Primed condition; there was no difference in the Vowel-not-Primed condition. The mean power spectra for the vowel minimal pair, *still-stool*, measured at the temporal midpoint of the fricative spectra, and their respective differences for one of these participants are displayed in Figure 1. In addition, Kawamoto and colleagues found that the duration of the initial segment was significantly longer in the Vowel-Primed condition, where the initial segment immediately preceded the unknown segment, than in the Vowel-not-Primed condition, where the initial segment was a segment removed from the unknown segment.

Together, the findings of Kawamoto and colleagues (2007) and Whalen (1990) are fully consistent with the minimal unit of articulation being the segment for which anticipatory coarticulation effects are not a necessary and obligatory accompaniment of speech. At the same time, the findings of Kawamoto and colleagues (2007) showed that articulation could proceed on the basis of either the segment or whole word criteria. When it is done so on the basis of the segment criterion, anticipatory coarticulation effects are only produced when the relevant segments are known. When the criterion to initiate articulation is the whole word, anticipatory coarticulation effects are produced regardless of the priming condition.

A Connectionist Implementation Using a Sub-syllabic Minimal Unit of Articulation

In this section, we outline a connectionist network that models the articulation process in which the minimal unit of articulation is the segment. The network takes as input the output of the phonological encoding process (not modeled here) corresponding to the phonological code of a syllable (the "plan"), and generates a sequence of articulatory movements (see Figure 2). The criterion to initiate articulation is reflected by the nature of the input to the network — a complete syllabic code corresponds to the whole word criterion and an incomplete syllabic code that can be incrementally updated to reflect ongoing processing corresponds to the segment criterion. The goal of the current implementation is to demonstrate how the presence and absence of coarticulation effects as well

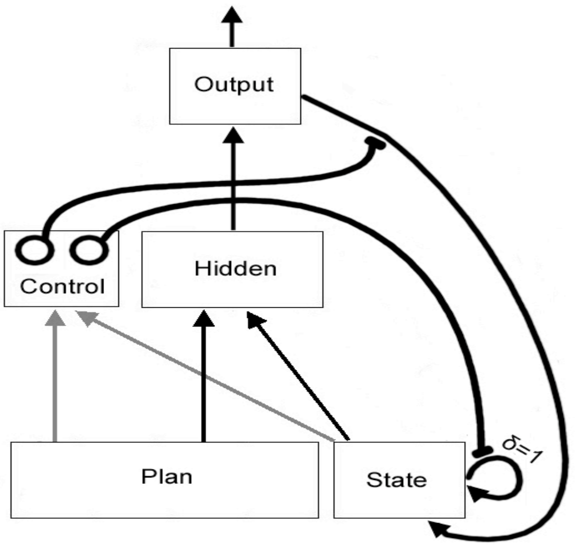

Figure 2: The basic architecture of the current model. The black arrows denote full connectivity with the exception that the connections between output-to-state and state-to-state are one-to-one. The gray arrows denote that only some inputs are fed into the next layer (see the discussion of the Control Structure below).

as local segment duration effects can be accounted for when the segment criterion is used.

The Representations Used

The input representation used here is a slot based local representation scheme that specifies the segmental content, syllabic frame, and encoding status of a metrical slot (see Table 1). The output and state representations correspond to the syllabic position of the current segment being produced and a small set of articulatory dimensions: the velar opening, the positions of the tongue tip and tongue body, the vertical and horizontal lip separations, and the constriction of the glottis (see Table 1).

Network Architecture

The network consists of two components — a Jordan net that produces a sequence of articulatory movements from a plan, and a control structure that controls how long the current segment is articulated. For the current purposes, the model was designed to simply produce one segment with each sweep, starting from the neutral state and returning to it when articulation is complete.

Jordan Net. A Jordan net was chosen to implement the articulatory pathway because its ability to simulate anticipatory coarticulation is well documented (e.g., Jordan, 1986). Similar approaches to model speech production using a Jordan network have previously been proposed (e.g., Dell, Juliano, Govindjee, 1993; Plaut & Kello, 1999). This component of the network contained three fully connected layers: The first contained the plan and state units; the second the hidden units; and the third the output units. However, the connections between output-to-state and state-to-state are one-to-one. In the current network, the decay parameter is set to equal 1 (i.e., δ=1; no decay). However, due to the actions of the Sigma-Pi connections from the control structure (see below), only the output of a single sweep will be buffered in the state units at any given time.

Control Structure. The control structure is a simple feed forward network that acts as a monitoring and gating mechanism that checks the segment currently being produced with what has to be produced next. If information about the following segment is available, the following segment will be produced in the next sweep. If not, the model will simply continue the articulation of the current segment until the next segment becomes available.

Table 1: Examples of the input and output representations used in the current model. Input for the control structure (in bold), with the "-" denoting the unspecified segment unit (used only for priming), and the "$" denoting the metrical slot unit (specified or not). The first and last sweep represents the neutral state.

	The Complete Input Plan for *stool*																									
	Onset1							Onset2						Vowel								Coda				
Sweep	s	p	t	r	l	**-**	**$**	p	t	r	l	-	$	a	e	i	u	I	U	**-**	**$**	p	t	l	**-**	**$**
1	0	0	0	0	0	**0**	**0**	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**	0	0	0	**0**	**0**
2-5	1	0	0	0	0	**0**	**1**	0	1	0	0	**0**	**1**	0	0	0	1	0	0	**0**	**1**	0	0	1	**0**	**1**
6	0	0	0	0	0	**0**	**0**	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**	0	0	0	**0**	**0**

The Corresponding Target Output Values										
Sweep	Current Segment				Velar Opening	Tongue Tip	Tongue Body	Vertical Lip Separation	Horizontal Lip Separation	Glottal Opening
1	0	0	0	0	.9	.5	.2	.6	.5	.8
2	1	0	0	0	.1	.8	.85	.8	.15	.8
3	0	1	0	0	.1	.9	.9	.83	.05	.8
4	0	0	1	0	.1	.1	.1	.85	.01	.1
5	0	0	0	1	.1	.4	.17	.35	.3	.1
6	0	0	0	0	.9	.5	.2	.6	.5	.8

This control is accomplished by taking as inputs the unspecified segment units and the metrical slot units (i.e., the “-“ and “$” units in Table 1) from the plan unit and the state units that buffer the output of the current segment units. The unspecified segment units denote whether or not the identity of a particular segment is unknown (1=”on” or unknown), the metrical slot units denote whether or not a particular metrical slot is specified in the word frame (1=”on” or specified), and the current segment units denote the syllabic position of the segment produced in the previous sweep. Together, these units feed into 2 control units, one for the Sigma-Pi connections to the output-to-state connections and the other for the Sigma-Pi connections to the state-to-state connections, and turn them “on” or “off” accordingly.

Specifically, the output-to-state connections will be turned “off” and the state-to-state connections turned “on” if the input to the unspecified segment unit and the metrical slot unit that immediately follows the segment produced in the previous sweep are both “on”. Otherwise, the output-to-state connections will be “on” and the state-to-state connections will be “off” (see Table 2). In the former, the model will behave as a feed forward network that will produce either the same or a modified version (e.g., if the incomplete plan is updated incrementally from *s___l→ s_ool→stool*, the articulation of the /s/ will transition from uncoarticulated to fully coarticulated when the vowel becomes known) of the articulatory output from an earlier sweep. In the latter, the model will behave as a Jordan net with only the previous time step buffered.

Table 2: The input-output correspondence for the control units. The syllabic position of the unspecified segment, encoding status, and current segment from left to right are Onset1, Onset2, Vowel, and Coda. Only the relevant values are displayed here.

From the Input Plan		From State	Control Units	
Unspecified Segment	Metrical Slot	Current Segment	State-to -State	Output-to-State
1 * * *	1 * * *	0 0 0 0	1	0
* 1 * *	* 1 * *	1 0 0 0	1	0
* * 1 *	* * 1 *	0 1 0 0	1	0
* * * 1	* * * 1	0 0 1 0	1	0
All other circumstances			0	1

Training Network

The Jordan net was trained on the Tlearn simulator (Plunkett & Elman, 1997). The training corpus for the Jordan net consisted of the input and output sequences corresponding to the words *sip*, *soup*, *sit*, *suit*, *spill*, *spool*, *still*, and *stool*. Training was carried out for 3000 epochs, with a learning rate = 0.05 and momentum = 0.

Although the behavior of the control structure is that of a simple table lookup and can be hardwired, the control structure was trained in the simulations described below to show that the relationship between what is being produced and the availability of what has to be produced next can be learned and used accordingly. Training of the control structure was carried out separately on a representative set of input and output sequences that covered the range of possible circumstances that can arise in the experiment (described above) carried out by Kawamoto et al. (2007). The network was trained for 2000 epochs, with a learning rate = 1 and momentum = 0.

Testing the Network

Simulation of the Vowel-Primed and Vowel-not-Primed results for the whole word and segment criteria was carried out. For the whole word criterion, input to the network corresponds to the complete plan from the outset (i.e., the training input sequences) for both priming conditions. However, for the segment criterion, the initial plans in the test sequences included only those segments that were known beforehand (i.e., *s_ool*, *s_ill*, *st__l*, or *sp__l*) with the unspecified segment represented by the “-“ unit in the appropriate metrical slot. The missing segment was added later for each target (i.e., *spill*, *spool*, *still*, or *stool*). In all, 8 novel test sequences were used: 4 corresponding to the Vowel-Primed sequences (e.g., *s_ool→spool*, *s_ool→stool*) and 4 corresponding to the Vowel-not-Primed sequences (e.g., *st__l→stool* and *st__l→still*). For these input sequences, the changing input changes the status of the control units.

To demonstrate the presence and absence of coarticulatory effects as well as local segment duration effects, the outputs of the network were computed offline from the weight matrices of the Jordan net and the control structure. The 8 novel test sequences were lengthened to 10 sweeps beginning with the neutral state. An incomplete plan was then presented for 5 sweeps in the Vowel-Primed sequences and 6 sweeps in the Vowel-not-Primed sequences. Next the complete plan corresponding to one of the four targets was presented in the subsequent sweeps. Finally, the neutral state was presented again as the final sweep. Note that the number of sweeps that the incomplete plan was presented in the novel test sequences was simply to show that articulation could be initiated long before the complete phonological code is available.

Coarticulatory Effects. For the current demonstration in which the spread and rounded vowels are being articulated, the articulatory dimensions of interest are the vertical and horizontal lip separations. The lip configurations for the word *stool* produced with the complete training input sequence and the novel test sequences *s_ool→stool* and *st__l→stool* in which the input is initially not a complete syllabic representation are displayed in Figure 3. The results are very clear. The outputs of the novel test sequence *s_ool→stool* are almost identical to the outputs corresponding to the training input sequence for *stool*. For both sequences, the /s/es were produced with a large vertical lip separation and a small

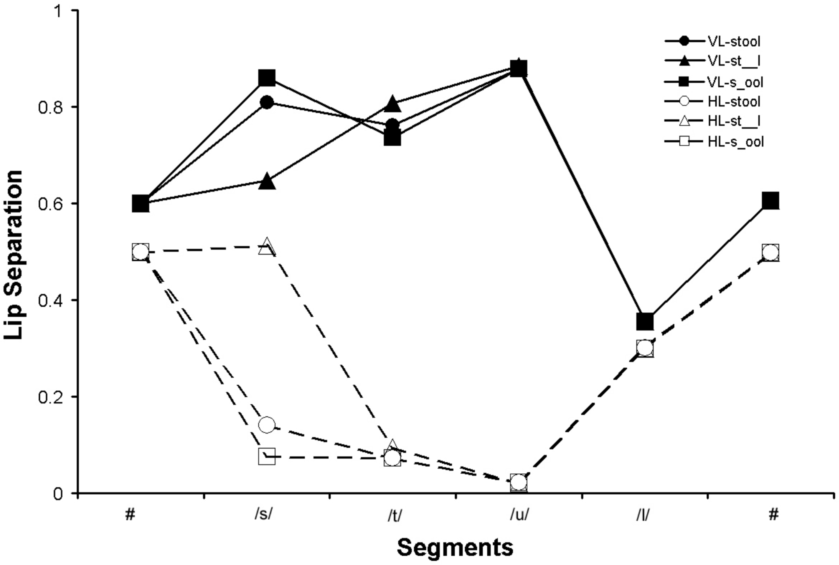

Figure 3: The Vertical and Horizontal lip separations (VL and HL) for the output corresponding to the complete plan *stool* as well as the novel input sequences *s_ool→stool* and *st__l→stool* (*s_ool* and *st__l* for short). The values corresponding to # denotes the lip separations at neutral state.

horizontal lip separation that corresponded to the pursing of the lips associated with rounding. However, the lip configuration during the production of the initial /s/ for the novel test sequence *st__l→stool* varied little from the neutral state (a lip configuration that is slight parted, but neither pursed nor spread) and was markedly different from the lip configurations of both the novel sequence *s_ool→stool* and the training sequence for *stool*.

Similar results with respect to lip spreading in the /s/ emerged with the training sequence for *still* and the novel test sequences *s_ill→still* and *st__l→still* (not displayed). The overall pattern of coarticulatory effects for the training sequences for *stool* and *still* and their 4 respective novel test sequences (described above) was also observed for the training sequences for *spool* and *spill* and their respective novel test sequences.

Local Segment Duration Effects. The results for the novel Vowel-Primed sequence *s_ool→stool*, showed that the articulation of the segment immediately preceding the unknown segment is lengthened until it becomes available (Table 3). Specifically, in sweep 2, the network produces the articulatory parameters for initial segment specified in the plan (i.e., /s/) and denotes that the segment being produced is the initial segment (the current segment units). However, in sweep 3, the input to the control units

Table 3: The novel test sequence *s_ool→stool* (i.e., the Vowel-Primed condition) and the corresponding output sequence as well as the actions of the control units at each sweep are displayed. Information for the 2nd segment was updated in sweep 7 (in bold) to correspond to the target *stool.*

	The Vowel Prime input example for *s_ool→stool*																										
	Onset1							Onset2						Vowel								Coda1				
Sweep	s	p	t	r	l	-	$	p	t	r	l	-	$	a	e	i	u	I	U	-	$	p	t	l	-	$
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2-6	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0	1
7-9	1	0	0	0	0	0	1	**0**	**1**	**0**	**0**	**0**	**1**	0	0	0	1	0	0	0	1	0	0	1	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	The Corresponding Output Values										Control Units	
Sweep	Current Segment				Velar Opening	Tongue Tip	Tongue Body	Vertical Lip Separation	Horizontal Lip Separation	Glottal Opening	State-State	Output-State
1	.00	.00	.00	.00	.90	.50	.20	.60	.50	.80	.00	1.00
2	.97	.00	.01	.01	.07	.73	.85	.86	.08	.79	.00	1.00
3-6	.97	.00	.01	.01	.07	.73	.85	.86	.08	.79	1.00	.00
7	.00	.98	.03	.00	.09	.75	.32	.74	.07	.79	.00	1.00
8	.01	.01	.99	.02	.11	.13	.12	.88	.02	.14	.00	1.00
9	.01	.00	.01	.99	.11	.46	.14	.36	.30	.14	.00	1.00
10	.00	.00	.00	.01	.90	.48	.19	.61	.50	.79	.00	1.00

indicates that what has to be produced next is still unknown, and thus the control units turn “on” the state-to-state connections and turn “off“ the output-to-state connections. These actions turn the Jordan net into a feed forward network that simply updates the output of the earlier sweep. In essence, the articulation of the initial segment is lengthened until sweep 7, when the 2nd segment becomes known. At that point, control units turn the network back into the Jordan net and produced 2nd segment in sweep 7 and the remaining segments in the subsequent sweeps.

Conclusion

The results of the current simulations offer a succinct demonstration that a single network using a sub-syllabic minimal unit can account for both the presence and absence of anticipatory coarticulation effects as well as local segment duration effects. This approach can easily be extended to other local segment duration effects that can arise from a variety of processing difficulties in speech production. Moreover, the current network can be coupled to existing models, such as that of Dell et al. (1993), to account for a wider range of empirical data (e.g., latency data). Such an extension provides a way to explore the interplay between different dependent measures such as latency and segment duration.

Acknowledgments

This work was partially supported by a grant from the UCSC faculty senate. We thank the four anonymous reviewers for their comments and suggestions on an earlier version.

References

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Sciences, 17,* 149-195.

Fujimura, O. & Lovins, J. B. (1978) Syllables as concatenative phonetic units. In A. Bell & J. B. Hooper (Eds.), *Syllables and segments*, Amsterdam: North-Holland.

Jordan, M. (1986). Serial order: A parallel distributed processing approach. Technical Report 8604. San Diego: Institute for Cognitive Science. University of California.

Kawamoto, A. H. (1999). Incremental encoding and incremental articulation in speech production: Evidence based on response latency and initial segment duration. *Behavioral and Brain Sciences, 22*, 48-49.

Kawamoto, A. H., Kello, C. T., Jones, R. M., & Bame, K. (1998). Initial phoneme versus whole word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, & Cognition. 24*, 862-885.

Kawamoto, A. H., Liu, Q., Kehrlein, K., & Johnson, G. (2007). Fine-Grained Analysis of the Time-Course of Single Word Utterances: Testing Assumptions in Speech Production and Reading Aloud. Unpublished manuscript.

Kello, C, T. (2004). Control over the time course of cognition in the Tempo-Naming task. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 942-955.

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

Mackay, D. G. (1987). The organization of perception and action: A theory for language and other cognitive skills. New York: Springer-Verlag.

Meyer, A. S., Roelofs, A., & Levelt, W. J. M. (2003), Word length effect in object naming: The Role of a response criterion. *Journal of Memory and Language, 48*, 131-147.

Plunkett, K. & Elman, J. L. (1997) Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations. Cambridge, MA: MIT Press

Plaut, D. C. and Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381-415). Mahwah, NJ: Erlbaum.

Rastle, K., Harrington, J., Coltheart, M., & Palethorpe, S. (2000). Reading aloud begins when the computation of phonology is complete. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1178-1191.

Rumelhart, D. E., Hinton, G., & McClelland, J. L. (1986), A general framework for parallel distributed processing. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel Distributed Processing: Exploration in the Microstructure of Cognition: Vol. 1. Foundations* (pp. 318-362). Cambridge, MA:MIT Press.

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics, 18*, 3-35.