AN ANALYSIS OF THE RELATIVE TIMING OF COARTICULATED GESTURES WITHIN VCV SEQUENCES

Juraj Šimko[a], *Fred Cummins*[b] & *Štefan Beňuš*[c,d]

[a]CITEC, Bielefeld University, Germany; [b]University College Dublin, Ireland;
[c]Constantine the Philosopher University, Nitra, Slovakia;
[d]Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
juraj.simko@uni-bielefeld.de; fred.cummins@ucd.ie; sbenus@ukf.sk

ABSTRACT

Accounting for the fine details of the patterning of co-articulated gestures in time is an important outstanding challenge. We report articulatory data capturing the relative timing of consonantal and final vowel gestures in VCV sequences. The elicitation procedure used ensures that the speech studied exhibits substantial variation in both speaking rate and in the degree of hypo/hyper-articulation. We find that both gradient and categorical effects appear to underlie the resulting inter-gestural timing. The results are compatible with a view of coordination based on efficiency principles.

Keywords: coarticulation, gestural timing, coordination.

1. INTRODUCTION

Accounting for the temporal details of sequences of discrete actions remains a central challenge to understanding behavior [5]. This concern arises with redoubled force in speech production, where articulatory gestures are coordinated along multiple parallel tiers. Shortcomings in extant theories of gestural sequencing are manifested, e.g., in state of the art articulatory synthesis, where principles for coordinating such gestural streams are still outstanding. For example, the influential neural model of speech production, DIVA, simply presumes that an articulatory movement is triggered at the moment when the previous movement successfully reaches its target [4].

Empirical acoustic and articulatory studies have long shown that this naive approach to sequencing does not reflect the way speech gestures are organized in time by human speakers. In his acoustic analysis of simple VCV sequences, Öhman [8] has shown that the transition of the vocal tract towards the second vowel starts well before acoustic closure for the medial stop consonant is achieved. More recent articulatory analysis of gestural timing in VCV sequences with a bilabial consonantal gesture by Löfqvist and Gracco [7] has demonstrated that the onset of the tongue body movement towards the vowel following the bilabial stop may actually *precede* the onset of lip movement towards the closure. In particular, this seemed to be the prevalent sequencing strategy for productions of /ipa/ and /iba/, while in the reversed vocalic context (/abi/, /api/) the lip closing gesture started before the onset of the transition towards the second vowel.

These results suggest some important general principles that any account of gestural sequencing must respect, and they raise some questions. From both studies, it is clear that an individual gesture is not triggered at some fixed point in the unfolding trajectory of the preceding gesture. Indeed, phonetic analysis provides some support for the phonological postulate that vowel sequences are relatively independent of intervening consonantal gestures [3]. Further evidence for the phonological postulate comes, for example, from processes of vowel harmony which are blind to inter-vowel consonants, while it finds direct theoretical representation in the separation of vowel and consonantal tiers in autosegmental representations [2]. On this view, speech production might be first considered as the articulation of a series of vowels (acting as syllabic nuclei), and consonantal gestures can then be understood as context-specific addenda to this sequence. If we adopt this position, then we can seek to understand the timing of the consonantal gesture in VCV sequences *relative to the onset of the second vowel*. The Löfqvist and Gracco data do not provide sufficient information to tell whether the difference in serial order observed between /iba/ and /abi/ is best understood as a categorical difference associated with different vowel phonemes, or as a gradient difference related to the interplay between the specific articulatory demands of the vowels and the consonant.

This context sensitivity is not addressed in most theories of gestural sequencing, although the need

to provide a fuller account of the context-sensitive interleaving of gestures in speech is keenly felt. A notable exception is the theory based on optimality principles and embodiment of speech action recently presented by Simko and Cummins [10, 11] inspired by Lindblom's work on Hypo- and Hyper-articulation [6], and based on an extension to the theories of articulatory phonology [1] and its task dynamic implementation [9]. We assume that gestures are sequentially organized in an optimal manner that reflects both the physiological characteristics of the vocal tract and the demands on parsing the resulting sequence by a listener. The implementation of this optimization theory scores candidate gestural sequences using a cost function comprising a weighted combination of articulatory effort (a speaker-based cost), ease of parsing (listener-based), and the premium put on speaking rate (situation-based). They showed that the sequences /abi/ and /iba/ that minimize this cost function reproduce the qualitative aspects of the above mentioned sequencing details as reported by Löfqvist and Gracco [7]. In an optimal sequence, the bilabial gesture is triggered at the appropriate time and with the appropriate force so as to guarantee the realization of the stop between the flanking vowels, in a manner that is readily perceivable by the listener. Moreover, the timing and force are efficient in the given context: In the transit from /i/ to /a/ in /iba/, the initial relatively high position of the tongue and jaw means that the lip aperture is smaller and the consonantal gesture can thus start later, or be effected with less force, than in the complementary sequence /abi/.

The efficiency considerations that underlie our model suggest that the variation in the relative timing of consonant and postconsonantal vowel gestures might be lawfully related to either articulatory distance and/or articulatory force, and that these, in turn, will depend on the degree to which the utterance is hyper- or hypo-articulated, and the rate at which it is spoken. We here present an initial analysis of articulatory data obtained in a manner that ensures rich variation in both dimensions. This allows us to test the hypothesis based on efficiency considerations that the relative timing of the two gestures co-vary with initial displacement and/or deployment force of the consonantal gesture.

We test this prediction against the background of an alternative account consistent with the measurements of Löfqvist and Gracco, namely that the difference between the timing patterns in /abi/ and /iba/ sequences reflects phonological systematicity, i.e., is linked to the phonological contrast between the underlying vocalic contexts.

2. DATA COLLECTION AND PROCESSING

We used electro-magnetic articulography (EMA) to track the movements of receivers attached to active articulators at a sampling rate of 200 Hz. We here report an analysis of the first principal component of movement of the two lips and the tongue body. Four subjects read meaningful Slovak sentences containing real words *iba* 'only' and *abi* 'in order to'. These target VCV sequences were flanked by bilabial nasals /m/ and contrasting vowels. Thus, the speakers produced embedded sequences /…am\#iba\#mu…/ and /…im\#abi\#mu…/.

Recordings for each VCV sequence were done under two conditions. First, rate was varied by raising/lowering experimenter's hand as a prompt to speak more/less rapidly. In the second condition the same hand signals were used to vary the degree of hyper-articulation produced, thus encouraging the subject to range from very lax, indistinct productions to very clear, hyper-articulated ones. Using this method, between 135 and 237 readings were recorded for each speaker and each VCV sequence.

2.1. Labeling

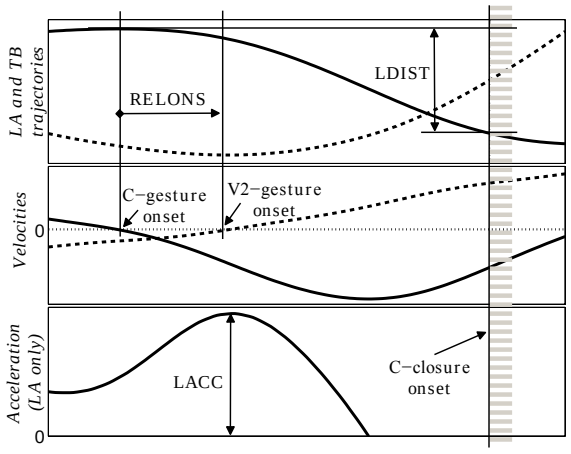
A single annotator identified the onset and offset of bilabial closure for /b/ in the acoustic signal. Given the large purposeful variation in tempo and precision of the prompt sentences, great variability of closure types occurred creating a continuum between clear stop-like closures with the complete absence of formants and a clear burst after the release, through short, often nasalized, closures but with discontinuous changes in the formant trajectories and waveform, to incomplete closures of bilabial /w/-like approximants with minimal and continuous changes in the signal. Only the tokens with reliable acoustic markers of a closure, i.e. the first two cases above, are considered in this paper. The number of analyzed samples for each subject and each VCV sequence thus varied considerably, between 18 /iba/ tokens of subject S4 to 110 of /abi/ recordings of subject S2.

A lip aperture measure was calculated as the Euclidean distance between the positions of the lower lip and upper lip sensors. The onset of lip closure movement for the bilabial stop /b/ was

identified as a velocity zero-crossing of the principal component of the derived lip aperture signal. Similarly, the onset of the tongue body movement towards the vowel following /b/ was placed at the velocity zero-crossing of the principal component of sensor TB2 signal.

To approximate the distance between the lips at the onset of the bilabial gesture, LDIST (lip distance) was defined as the difference between the value of lip aperture at the onset of the consonantal gesture and at the onset of the acoustically defined bilabial closure. The relative timing of the onsets of the consonantal and vowel gestures is described by RELONS (relative onset), and is computed as the time at which the tongue begins to move towards the second vowel, minus the time at which the lip closing gesture begins. Positive values thus mean that the consonantal gesture starts *before* the vocalic gesture. Finally, LACC (lip acceleration) stands for the peak acceleration of the lip aperture movement during the interval from the movement onset to the peak velocity of the closure movement. As acceleration is lawfully related to force, we shall use this measure as an estimate of the force driving the lips towards each other. These three variables are illustrated in Fig. 1.

Figure 1: Position, velocity and acceleration of lip aperture (solid lines) and tongue body (dashed lines, top 2 panels only). The variables RELONS, LDIST and LACC are described in the text.



3. RESULTS

To establish whether gestural timing and dynamics depend on segmental identity, we report mean values in Table 1. The mean relative onset of C and V2 gestures differ substantially from those reported in Löfqvist and Gracco [7] in that for three of four speakers, the lip gesture starts relatively earlier in /iba/ than in /abi/. Only subject S4 displayed the pattern reported by Löfqvist and Gracco where tongue onset is earlier than lip gesture onset in /iba/, and the reverse for /abi/. Interestingly, there is little difference in the average size of the C gesture (LDIST) for the same three speakers, although the difference, while small, is significant for S3. This might be attributable to an effect on the lip movement of the bilabial nasal preceding the analyzed VCV sequences. On the other hand, the mean acceleration, which we interpret as a proxy for force, is significantly different across the two word classes for all subjects.

Mean values are useful for testing whether the phoneme identity alone is useful in predicting articulatory detail. A notable feature of our database is the amount of variation elicited in both rate and in the degree of hyper- and hypo-articulation. We therefore next examine the degree to which both LDIST and LACC might covary with RELONS.

Table 1: Mean relative onset, lip aperture at onset, and consonantal acceleration. Units are ms, mm, and ms^{-2} respectively. The *p*-values are based on two-tailed *t*-tests. **: $p<.01$,***: $p<.001$.

	RELONS		LDIST		LACC	
	/abi/	/iba/	/abi/	/iba/	/abi/	/iba/
S1	-6.3	7.6	3.2	3.0	12.2	10.1
	$t = 6.35$***		$t = 0.29$ n.s.		$t = 4.29$***	
S2	3.7	17.1	5.6	5.5	8.7	6.9
	$t = 6.25$***		$t = 0.26$ n.s.		$t = 6.48$***	
S3	6.4	11.3	4.8	3.6	6.8	4.3
	$t = 2.77$**		$t = 2.76$**		$t = 7.28$***	
S4	31.4	6.3	12.0	5.0	6.2	4.7
	$t = 5.34$***		$t = 5.57$***		$t = 3.32$**	

Table 2 provides results of single-variable linear regressions of LDIST and LACC on RELONS (columns 2-3 and 4-5, respectively) and a two-variable linear regression of both LDIST and LACC on RELONS (columns 6-8). Both LDIST and LACC were log-transformed, as this ameliorated the heteroscedasticity and non-normality of errors. The slope of the regression, its significance and the proportion of variance accounted for are provided.

The articulatory distance covered (LDIST), which was not different, on average, across the two syllable types, is an effective predictor of the relative timing of the consonantal onset with respect to the vowel gesture. The slopes are significantly positive for all fitted data sets with an exception of /iba/ sequences for speaker S4. The amount of variation this variable alone accounts for ranges from 0.56 to 0.11, and is considerably lower for /iba/ se-

quences than for /abi/. This suggests that both gradient processes (articulatory distance) and phonological identity (/iba/ vs /abi/) are at play here. Phonological identity alone was associated with a difference in mean acceleration (and hence force), but LACC is a very weak predictor of relative timing when considered as a continuous variable.

A composite model using both LDIST and LACC as explanatory variables suggests complementary influences of initial distance and peak acceleration on the value of RELONS: all significant regression slopes of LDIST variable are positive while all significant slopes of LACC are negative.

Table 2: Single predictor models predicting RELONS using either LDIST or LACC and the combined model. Figures shown are the slope of the best linear fit, and the adjusted proportion of variance accounted for by the regression. The significance of slope signs: *: $p<.05$, **: $p<.01$,***: $p<.001$.

	LDIST		LACC		LDIST	LACC	
	slope	R^2	slope	R^2	slope	slope	R^2
S1 /abi/	13.1***	0.42	10.5 n.s.	0.01	13.0***	8.27 n.s	0.43
/iba/	4.9**	0.11	-8.7*	0.08	5.6***	-40.5***	0.22
together	9.5***	0.21	-10.4**	0.04	9.9***	-12.1***	0.28
S2 /abi/	10.9***	0.56	-4.6 n.s.	0.01	10.9***	-5.3*	0.58
/iba/	32.1***	0.47	-28.7***	0.20	29.1***	-20.8***	0.57
together	18.6***	0.36	-25.4***	0.22	17.2***	-22.4***	0.51
S3 /abi/	10.4***	0.31	-7.3**	0.10	11.9***	**9.8****	0.51
/iba/	9.6***	0.21	-8.5 n.s.	0.03	9.8***	-9.1*	0.26
together	7.6***	0.15	-8.7***	0.11	9.8***	-12.0***	0.36
S4 /abi/	24.4***	0.53	21.8*	0.10	37.4**	-31.8**	0.63
/iba/	-11.9 n.s.	0.15	-17.5 n.s.	0.10	-10.4 n.s.	-14.5 n.s.	0.21
together	21.9***	0.45	23.5**	0.12	27.0***	-14.6 n.s.	0.47

4. DISCUSSION

The two predictors considered seem to play different roles. Whereas initial lip distance covaries with the magnitude of the interval between C onset and V2 onset, irrespective of phonological identity, peak velocity (force) displays variation that seems to be more closely linked to the categorical distinction between /iba/ and /abi/. When the effect of vocalic context on initial lip displacement is minimal (as for subjects S1--S3), the results of the categorical analysis indicate that greater articulatory forces are associated with later onsets of the bilabial gesture.

The positive regression slopes of LDIST suggest a link between the timing of consonantal gesture and state of articulators at the onset of the gesture. This dependency makes sense, and is compatible with our understanding of the role of efficiency considerations in determining the form of coarticulation: to traverse a larger distance, the consonantal gesture may start earlier.

The significantly negative slopes of LACC, which is a proxy for force, provide independent support for an efficiency-based interpretation, but the effect is weak, and LACC is not a good continuous predictor of the relative timing of the gestures.

5. REFERENCES

[1] Browman, C.P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49,155-180.

[2] Clements, G.N. 1993. The internal organization of speech sounds. In Goldsmith, J. (ed.), *A Handbook in Phonological Theory*. Blackwell.

[3] Fowler, C.A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in sequences of monosyllabic stress feet. *J. of Exp. Psychology: General* 112, 386-412.

[4] Guenther, F.H. 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review* 102(3), 594-621.

[5] Lashley, K.S. 1951. The problem of serial order in behavior. In Jeffress, L.A. (ed.), *Cerebral Mechanisms in Behavior* Wiley, 112-136.

[6] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H Theory. In Hardcastle, W.J., Marchal, A. (eds.), *Speech Production and Speech Modelling* Kluwer Academic Publishers, 403-439.

[7] Löfqvist, A., Gracco, V.L. 1999. Interarticulator programming in VCV sequences: Lip and tongue movements. *JASA* 105, 1864-1876.

[8] Öhman, S.E.G. 1966. Coarticulation in VCV Utterances: Spectrographic measurements. *JASA* 39(1), 151-168.

[9] Saltzman, E.L., Munhall, K.G. 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology* 1(4), 333-382.

[10] Simko, J., Cummins, F. 2010. Embodied task dynamics. *Psychological Review* 117(4), 1229-1246.

[11] Simko, J., Cummins, F. 2011. Sequencing and optimization within an embodied task dynamic model. *Cognitive Science* 35(3), 527-562.