Investigating the stability of intergestural timing relations

Juraj Šimko[1], *Fred Cummins*[2], *Štefan Beňuš*[3,4]

[1]CITEC, Bielefeld University, Germany, [2]University College Dublin, Ireland
[3]Slovak Academy of Sciences, Bratislava, Slovakia
[4]Constantine the Philosopher University, Nitra, Slovakia

juraj.simko@uni-bielefeld.de fred.cummins@ucd.ie sbenus@ukf.sk

Abstract

An articulatory analysis of lip and tongue coordination in VCV sequences is presented for four Slovak speakers. Lip and tongue movements are obtained for many utterances elicited in a manner that ensures great variation in both rate and in articulatory precision. Theory and models suggest that gestures might not be sequenced in simple linear order, but that medial consonant timing may be tied to the evolution of the following vowel gesture. We find that the relative timing of the consonant and second vowel gestures is most stable and exhibits least variability when the second vowel gesture provides the temporal reference frame. This work contributes to our understanding of non-linear coarticulatory effects in continuous, and variable, speech.

Index Terms: articulatory synthesis, intergestural timing, phase, coordination

1. Introduction

State of the art articulatory synthesis presents an increasingly viable alternative to more traditional concatenative speech synthesis approaches. As it promises to provide means for eliciting physiologically plausible variation in voice quality, speaking rate, fundamental frequency and intensity, the articulatory synthesis paradigm is becoming a technology of note in the quest towards synthesizing prosodically rich, expressive speech.

In order to successfully replicate human-like speech abilities, however, we need to better understand how speech movements—gestures—are coordinated in space and, equally importantly, in time. We here present a preliminary analysis of articulatory and acoustic recordings of VCV sequences with the aim of identifying possible invariance relationships in the relative timing of speech gestures. By its nature, this investigation is two-fold. First, we need to establish a plausible temporal frame of reference for each individual gesture, i.e., another gesture or group of gestures with respect to which the gesture is timed. We can then try to identify articulatory or acoustic temporal landmarks—both in the reference frame and in the individual gesture—that might serve as anchor points for their mutual coordination. Most recent systems make use of the simplistic assumption that one articulatory movement is triggered at the moment when the previous movement reaches its target. They thus use the preceding articulatory movement as the frame of reference for the timing of a gesture. The presumed anchor points to be aligned are then the time when the vocal tract reaches a first target and the onset of the following gesture itself.

As empirical acoustic and articulatory studies have long shown, this naïve approach to sequencing does not accurately characterize the way speech gestures are actually organized in time. E.g., Öhman [1] demonstrated that the transition of the vocal tract towards the second vowel (V2) in VCV sequences may start well before acoustic closure for the medial stop consonant is achieved. More recently, Lofqvist and Gracco [2] used articulatory analysis of VCV sequences with a bilabial consonant to show 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. The order in which the articulatory movements are triggered thus does not simply reflect the linear sequence of the associated acoustic segments.

From these studies it is clear that the preceding gesture is not a universally suitable timing reference for a gesture. In a VCV context, as above, the timing of a bilabial consonant might be better understood as occurring relative to a comparatively independent vowel sequence. The view that speech production might be viewed as the articulation of a series of vowels (acting as syllabic nuclei), with consonantal gestures as context-specific addenda to this sequence, is not new [3]. After all, its direct theoretical representation is the separation of vowel and consonantal tiers in autosegmental phonology. Our analysis, both theoretical, in [4], and empirical herein, also provides tentative support to the claim that the onset of the consonantal gesture is timed with respect to the inter-vocalic transition, i.e., relative to the onset of the vowel *following* the consonant, and not the other way round as the naïve approach suggests.

Lofqvist and Gracco [2] showed that the relative time of the onset of the bilabial gesture depends sensitively on the flanking vowel identities. In the sequences /iba/ and /ipa/, the tongue movement to the second vowel began in advance of the onset of the lip closing gesture. In /abi/ and /api/, by contrast, the consonant gesture started before the movement towards the second vowel. Thus the *onset* of the consonantal gesture is too variable to serve as an anchor point in VCV timing.

Within Articulatory Phonology [5, 6], individual speech gestures are realized as low dimensional dynamical systems. Each gesture is implemented as a critically damped second-order system, and is coordinated with surrounding gestures through reference to the dynamical state, or phase, of each gesture. Although the gestures are implemented as damped systems, a temporal framework for each is established by considering an abstract, underlying, cycle of the undamped system. A salient event can be described as happening at a specific *phase* of the cycle of this abstract oscillator. Relative timing among gestures can be described by associating a salient point or phase in one cycle with a corresponding phase in the other.

Simko and Cummins [7, 4] have recently presented an extension of the task dynamic implementation of Articulatory Phonology. Their model is inspired by Lindblom's work on Hypo- and Hyper-articulation and Emergent Phonology [8, 9]. It derives fully specified temporal relations among gestures based on optimality principles. In their model, dynamical gestures are sequentially coordinated in an optimal way, where 'optimality' is operationalized by taking into account the physiological characteristics of the vocal tract, the demands on parsing the resulting sequence by a listener and the requirements imposed by environmental factors. Optimal VCV sequences generated by their model successfully reproduced qualitative aspect of the sequencing patterns of Lofqvist and Gracco [2] reported above. In [4], they examined a number of optimal VCV sequences that arose by simulating variations in speaking rate and in the degree of articulatory precision, as well as in the vocalic context (/a-i/ vs. /i-a/). They showed that the most stable point of alignment across all of the highly varied sequences was between the time at which consonantal closure was reached, on the one hand, and a specific phase (not the onset) of the following vowel, on the other. The finding that the coordination between C and the following V was more stable than any other candidate phase relation is of interest, not least because of the centrality of CV structure in the world's languages, its early appearance in infant babbling forms, and its perceived unmarked character.

Contrary to the assumptions of (early versions of) Articulatory Phonology, the closure achievement in the simulations could not be interpreted as an invariant phase of the consonantal gesture itself, as it varied strongly with the underlying vocalic context. That means that the emergent intergestural relation computed by the optimization model temporally aligns a dynamical event (phase) of the vocalic gesture with an articulatory/acoustic event (closure) of the concurrent consonant.

Simko and Cummins' modeling is to a large extent qualitative. That means that the predictions in [4] do not include precise phase values of the vowel cycle with which the consonantal gesture is aligned. Moreover, the embodied nature of their approach implicitly assumes the dependence of invariance relations (e.g. the anchor phase value) on the physiological parameters of a particular vocal tract, the phonological properties of the spoken language, the contingent learning history of an individual speaker, etc. It is thus to be expected that the invariance of the stable intergestural relations mentioned above is not generalizable across all speakers and languages. Rather, it may be one of several possible efficient strategies that a speaker can discover and adopt during her speech acquisition process. In this regard, speech would not differ from other forms of coordinated action, such as handwriting and gait, that display highly individual temporal patterning in skilled execution.

In this paper, we analyze articulatory traces from VCV syllables produced with a great deal of variation in both speech rate and in articulatory precision. On a subject-by-subject basis, we try to identify the best temporal reference frame for describing the relationship between the medial C and final V gestures, and any potential anchor point within that reference frame. First, we compute the phase values of the consonantal and vocalic abstract gestural cycles at which particular articulatory and acoustic events occur. Using variance tests, we compare the stability of intergestural phasing, using both the consonantal and vowel cycles as alternative reference frames. Finally, we investigate the effect of vocalic context (/a-i/ vs. /i-a/) on the stability of the selected phasing relations.

2. Data collection and processing

2.1. Articulatory data collection and labeling

We used electro-magnetic articulography (EMA) to track the movements of receivers attached to active articulators at a sampling rate of 200 Hz. 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 sequences /...am#iba#mu.../ and /...im#abi#mu.../. Readings for each VCV sequence were done under two conditions, designed to ensure a great deal of articulatory variability in the data. First, rate was varied by having an experimenter raise and lower his hand as a prompt to speak more/less rapidly. In the second condition the hand signals were used to vary the degree of hyper-articulation produced, 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.

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 tokens of subject S4 uttering sequence /iba/ to 110 /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 this lip aperture signal. The onset of the tongue body movement towards the vowel following the bilabial consonant was placed at the velocity zero-crossing of the first principal component of the two-dimensional trace (in the midsagital plane) obtained from a sensor on the tongue body. Peak velocity derivation was straightforward, and peak acceleration was identified in the interval starting 20 ms before the gesture's onset and ending at the peak velocity time.

2.2. Phase estimation

The phase of the gestural cycle is an abstract concept, and its estimation from real data relies on several assumptions. Sensor movement (absolute, for vocalic tongue body trajectory, or relative, for lip closure) is presumed to follow approximately the trajectory of a critically damped linear second-order dynamical system. Temporal landmarks in the behavior of such a system (when the system reaches its target, when it reaches maximal velocity and acceleration) are fully determined by the system's stiffness k. Through this constant, the system can be associated with its undamped equivalent, a harmonic oscillator of the same stiffness, that can act as a reference cycle. The harmonic oscillator oscillates periodically around its equilibrium, and its status at any given moment can be characterized by a single number ϕ called phase, i.e., an angle in the phase space of the dynamical system. Kelso et al. [10] argued that this phase can be used as a characterization of the state of the associated critically damped system, and hence can serve to specify an anchor linking two or more dynamical systems. This way of linking dynamical states is independent of absolute ("clock") time-intervals, and is thus invariant with respect to the rate and extent of the movement.

System stiffness is analytically linked to the duration of a

Table 1: Differences between median phases of various events in /abi/ and /iba/ sequences, measured with respect to the /b/ gestural cycle (columns 2–5) and the V2 cycle (columns 6–9), respectively, and p-values arising from a Mann-Whitney test for differences across vowel contexts.

	/b/ cycle phase			V2 cycle phase			
	V2 onset	V2 p. vel.	V2 p. acc.	/b/ onset	/b/ p. vel.	/b/ p. acc.	/b/ closure
S1	25.73	-46.60	-53.93	-18.2	17.89	16.04	17.86
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
S2	-21.13	-52.75	-77.04	-16.91	8.69	13.71	**1.15**
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.354
S3	-9.44	-46.12	-61.62	-6.67	**3.43**	**-0.24**	10.01
p-value	<0.001	<0.001	<0.001	<0.001	0.049	0.778	<0.001
S4	-35.96	**8.35**	**12.99**	32.25	**-7.90**	**-6.94**	**-2.27**
p-value	<0.001	0.150	0.117	<0.001	0.228	0.457	0.936

single cycle of the undamped oscillator $T = 2\pi/\sqrt{k}$. In time T the phase value completes the full circle of 360°. Any event occurring at time t after the onset of the cycle, i.e., the onset of the gesture, thus can be assigned a phase $\phi = 360\,t/T$ of the reference cycle of the undamped system.

Phase computed this way depends solely on the relative temporal position of the event with respect to the onset of the underlying gesture, and the gesture's dynamical parameter k. In this work, the stiffness k of a gesture was estimated as the ratio of peak velocity of the movement to the overall displacement during the movement. Using this estimate, we derived the relative timing of specific events associated with one gesture by describing them as phase values obtained with respect to the reference cycle of another gesture in VCV sequences.

3. Results

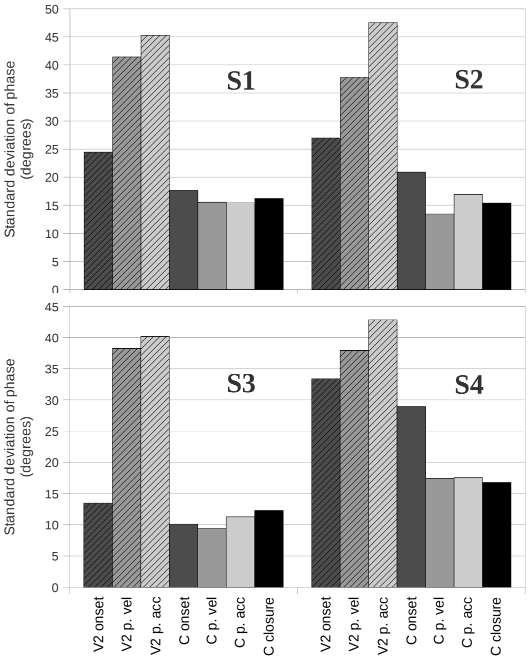
Fig. 1 shows standard deviations of phases at which various kinematic and acoustic events were achieved for all four subjects. For each subject, the standard deviations are computed from phases pooled over the entire set of recordings, irrespective of the speaking rate, articulatory precision and vocalic context. We identified three kinematic events for each analyzed speech gesture – its onset, the time when the movement reaches peak acceleration and the time of peak velocity – and the acoustic closure achievement for the consonantal gesture, and we computed the phase of its occurrence in the associated gesture—either the consonantal cycle or the final vowel. We didn't include the consonantal closure phase within the consonantal cycle, as it is not relevant for intergestural timing. Phases within the consonantal cycle are hatched. The light grey plain bars, for example, show the standard deviation of the phase of the vocalic cycle at which the consonantal gesture reached peak acceleration, while the hatched bars of the same shade of grey refer to peak accelerations of the vocalic gesture within the consonantal cycle.

A cursory look at the figure reveals that the standard deviation of phases expressed with respect to the vocalic cycle are considerably smaller than their consonantal cycle counterparts. This is true, irrespective of the specific kinematic landmark (gesture onset, peak velocity, etc) chosen. When broken apart by syllable type, the same pattern emerges, although the difference between the two reference cycles is somewhat more marked and consistent for /abi/ than for /iba/.

If smaller standard deviations imply a greater stability of intergestural timing alignment, this initial result can be interpreted as consistent with our theoretical expectations outlined in the introduction. In VCV sequences, the timing relationships are best expressed with reference to an underlying tier of vocalic gestures, and consonants are timed with respect to this tier. Fig. 1 thus suggests that a good candidate for an invariant coordination phase relation should be sought among the consonantal events expressed as phases of the V2 gestural cycle. The anchoring event should be robust with respect to continuous variations in speaking rate and articulatory precision, or degree of hypoarticulation, and should not be highly sensitive to the discrete changes of the underlying vocalic context (/a-i/ vs. /i-a/ in our recordings).

In order to rank the stability of the phases of kinematic and acoustic landmarks, we therefore investigated the effect of vo-

Figure 1: Standard deviations of phase values of various events measured with respect to the /b/ gestural cycle (hatched bars) and the V2 cycle (plain bars).



calic context on phase values of the same inter-gestural relations reported in Fig. 1. Table 1 lists the differences between median values of the phases in /iba/ sequences and median values of the phases in /abi/ sequences. For each landmark and reference cycle, we conducted a simple Mann-Whitney test for equality of underlying distributions, and the p values are provided in small script. For most candidate anchor points and reference cycles, there was a marked difference in median phase across the two vowel contexts. For gestural onsets (both consonantal onset as a phase of V2 cycle and V2-onset with respect to the consonant), the differences between the phase distributions are significant, suggesting that gestural onsets do not serve as anchor points in inter-gestural coordination. For the remaining landmarks, the differences between the medians are much more pronounced for /b/ cycle phase values than for V2 cycle ones. The influence of the vocalic context thus may be responsible for much of the difference in phase variance shown in Fig. 1.

Several phase relations exhibit relative stability with respect to vocalic context. For the onset of the consonantal closure measured as a phase of the V2 cycle, for Subjects 2 and 4, the difference between the medians is very small and the underlying distributions are not significantly different. The same is true for the peak acceleration of the consonantal gesture with respect to the V2 cycle for subject 3, and, to a lesser extent, for several other landmarks marked in bold font in the table.

4. Discussion

We set out to ask whether our data could provide hints about which events in the speech stream are temporally coordinated with respect to which temporal reference frames. Naïve sequencing models order phonemes, and perhaps gestures, in simple linear order, so that V2 would begin when the preceding C gesture reaches some landmark. Some empirical evidence [1, 2] has suggested that simple linear sequencing may not reflect what speakers do. Furthermore, both theoretical considerations [3], and our own modeling work [4] have suggested that consonants are timed with respect to an underlying sequence of vowels, i.e., that some point within the C gesture might be critically timed with respect to the vowels, and in particular, with respect to V2.

In examining the variability of phase relations found in our data, which, by design, exhibit very great variation in both rate and articulatory precision, we found that consonantal gesture timing with respect to V2 was markedly less variable than V2 gesture timing with respect to the consonant. This was consistent across all four speakers. The analysis of the influence of vocalic context also shows that the consonant timing with respect to the following vowel is less context sensitive that the timing of the vowel with respect to the preceding consonant.

On the other hand, it was not possible to identify a single anchor point that demonstrated phase invariance for all subjects. If gesture morphology is relatively constant, the very notion of specific anchor points may, indeed, be somewhat misleading. It is the relation of one gesture to another that we seek to characterize, and anchor points represent an operationalization that may be of help in determining stability, but that may not, in fact, privilege one specific landmark over another.

Speakers 2 and 4 seem to show CV coordination much in line with the predictions made by Simko and Cummins [4]. The most stable coordination found (low variance in phase and relatively invariant median phase with respect to the vocalic context) is the phasing of the consonantal closure with respect to the gestural cycle of the following vowel. For S4 in particular, the entire consonantal gesture appears to be relatively invariant when compared in time with the V2 gesture. It is worth noting that S4 also usually produced a clearly articulated phrase boundary immediately before the target VCV sequence, usually in the form of a glottal stop. This may have reduced the effect of coarticulatory phenomena arising outside the target sequence.

Our results are broadly consistent with phonologically sensible postulates of gestural sequencing, as well with some predictions based on the efficiency principles. As mentioned above, we have not succeeded in finding a coordination pattern fitting the behavior of all four analyzed speakers. The manner in which the phase value of an individual event has been estimated is based on several presumptions, e.g. the linear critically damped second-order dynamical nature of gestures. Approximating the articulatory trajectories with non-linear, subcritically damped dynamical systems of, perhaps, higher degree, may provide more complete data for the statistical analysis of temporal coordination. But it is also possible that there is no universal, speaker-independent answer to some of the questions stated in this work, and that each speaker selects one of several feasible and efficient temporal coordination strategies or a combination thereof.

Despite the open questions, we believe that we have outlined here a viable approach to addressing one of the most important questions in speech production studies, that of temporal coordination, that has direct relevance for building a flexible, expressive speech synthesis platform.

5. Acknowledgements

This work was partly supported by the EPSF project ITMS 26240220064 and VEGA project 2/0202/11 to the third author.

6. References

[1] S. E. G. Öhman, “Coarticulation in VCV Utterances: Spectrographic Measurements,” *Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[2] A. Löfqvist and V. L. Gracco, “Interarticulator programming in VCV sequences: Lip and tongue movements,” *Journal of the Acoustical Society of America*, vol. 105, pp. 1864–1876, 1999.

[3] C. A. Fowler, “Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in sequences of monosyllabic stress feet,” *Journal of Experimental Psychology: General*, vol. 112, pp. 386–412, 1983.

[4] J. Simko and F. Cummins, “Sequencing and optimization within an Embodied Task Dynamic model,” *Cognitive Science*, vol. 35, no. 3, pp. 527–562, 2011.

[5] C. P. Browman and L. Goldstein, “Articulatory Phonology: An Overview,” *Phonetica*, vol. 49, pp. 155–180, 1992.

[6] E. L. Saltzman and K. G. Munhall, “A Dynamical Approach to Gestural Patterning in Speech Production,” *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[7] J. Simko and F. Cummins, “Embodied Task Dynamics,” *Psychological Review*, vol. 117, no. 4, pp. 1229–1246, 2010.

[8] B. Lindblom, “Explaining Phonetic Variation: A Sketch of the H&H Theory,” in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Kluwer Academic Publishers, 1990, pp. 403–439.

[9] ——, “Emergent phonology,” in *Proc. 25th Annual Meeting of the Berkeley Linguistics Society*, U. California, Berkeley, 1999.

[10] J. A. S. Kelso, E. L. Saltzman, and B. Tuller, “The dynamical perspective on speech production: data and theory,” *Journal of Phonetics*, vol. 14, pp. 29–59, 1986.