Contents lists available at ScienceDirect

Journal of Phonetics

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

Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing

Štefan Beňuš [a,b,*], Juraj Šimko [c]

[a] Constantine the Philosopher University, Štefánikova 67, 949 74 Nitra, Slovakia
[b] Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava, Slovakia
[c] Bielefeld University, Univestitätsstraße 25, 33615 Bielefeld, Germany

ABSTRACT

The bulk of our current knowledge about articulatory/acoustic signatures of prosodic structure comes from paradigms that elicit discrete prosodic variation intentionally produced by subjects. In this paper, we collect speech elicited through continuous variation in tempo and hypo–hyper articulation, and analyze spontaneous emergence of high-level prosodic boundaries as a means of resolving low-level tempo and precision demands. Our data show that as the area of structural affordance for a prosodic boundary comes under decreasing temporal pressure, the temporal coordination patterns of the gestures in the vicinity of this affordance get continuously rearranged. This re-arrangement is comprehensively captured with the optimization-based embodied task dynamics platform (Šimko & Cummins, 2010, 2011), in which this phenomenon can be modeled in terms of localized changes in relative demands on articulatory efficiency, perceptual clarity, and minimal duration, and the optimal resolution of these demands.

1. Introduction

One of the fundamental linguistic functions of prosody is structuring speech by means of prominence distribution and unit segmentation to facilitate cognitive processing for both the speaker and the listener (Beckman, Edwards, & Fletcher, 1992; Byrd & Saltzman, 2003; Cho, 2006). The predominant research paradigm in analyzing prosodic structuring and effects of prosodic boundaries is to elicit functionally discrete variation assumed to arise either autonomously in the prosodic structure or through its interaction with the syntactic or information structures. For example, prosodic boundaries of at least two strengths were found to affect temporal aspects of speech in different ways, including pre-boundary and post-boundary strengthening and decrease of cross-boundary overlap (Byrd, 2000; Byrd, Krivokapić, & Lee, 2006; Byrd & Saltzman, 1998; Cho, 2004; Fougeron & Keating, 1997; Krivokapić, 2007). In this way, researchers look for invariant underlying principles that link functional linguistic contrasts with speech production.

This research program has been extremely fruitful. For example, the strength of prosodic boundaries – the core issue of the present paper – was found to correlate with the duration, stiffness and the degree of overlap of articulatory movements in the vicinity of these boundaries. In general, the stronger the boundary, the longer, less stiff and less overlapped the movements. Additionally, predictions of several proposed models correspond to many patterns observed in the data (Byrd & Saltzman, 2003; Cho, 2002).

The paradigm traditionally used for data elicitation focuses at discrete prosodic contrasts with high functional load. Speech elicited in this way invariably produces data that probe only limited intervals of the (theoretically possible) realizational continuum of the prosodic structure. Statistical effects of discrete contrasts – presence vs. absence of a break, or word boundary vs. intermediate (minor) boundary vs. intonational (major) boundary – provide snapshots of the ways prosodic structure is realized in these discontinuous intervals. However, data obtained using discrete elicitation methodologies might not be ideal for investigating dynamic and evolutionary aspects of relationship between boundary strength and articulation. Data covering prosodic continua, on the other hand, have potential for illuminating the dynamics underlying the prosodic structure and may help us understand why the observed patterns are stable and ultimately how they could have emerged. Furthermore, continuous data provide opportunity for pursuing one of the intended directions of research within articulatory modeling of speech mentioned by Byrd and Saltzman (2003), and explore the issue of gradient vs. categorical nature of articulatory signatures of prosodic structure related to the strength of a boundary.[1] This issue has been investigated by, for example, Krivokapić (2007) who tested predicted gradiency in both production and perception of boundary strength. Her data suggest that prosodic breaks are produced in a categorical fashion (two major categories; fewer than three or four predicted by phonological models)

[*] Corresponding author at: Constantine the Philosopher University, Štefánikova 67, 949 74 Nitra, Slovakia. Tel./fax: +421 37 6408455.
E-mail address: sbenus@ukf.sk (Š. Beňuš).

[1] Byrd and Saltzman (2003: 176) mention two kinds of gradiency: in terms of precise location of a break, and also in terms of the strength of a boundary. We are interested in this second notion of gradiency.

0095-4470/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2013.12.005

and perceived in a more gradient fashion (up to eight levels). Our goal is to further explore this issue by employing a different paradigm of data elicitation and analysis.

In addition to the issues of continuity in observed articulatory patterns and data elicitation, the focus of the current paper is on variation with respect to articulatory context. Most available studies vary prosody but not the context (Byrd, 2000; Byrd et al., 2006; Cho, 2004, 2006; see below for details). Consequently, the interplay between context, timing, and prosody is still relatively little understood.

Our aim is thus to investigate the articulatory patterns – their continuous change and contextual effects – in prosodically varied speech. To complement previous studies, our speech material includes variation of the prosodic boundary strength induced by low-level variations such as tempo and hypo–hyper articulation while attempting to control for the high-level functional prosodic contrasts commonly arising from syntactic, pragmatic or rhythmical structures. We believe that in this way we can draw insights into the role of prosody that are different, and complementary, to the knowledge gained from the studies outlined above.

1.1. Articulatory signatures of discrete prosodic breaks

Discrete prosodic contrasts assumed by the standard autosegmental-metrical model arise from the interface between prosody on the one hand and syntax or information structure on the other. Several studies reported the effects of the strength of prosodic boundaries on relative timing of consonant and vowel gestures in the vicinity of these boundaries (e.g. Beckman & Edwards, 1992; Byrd & Saltzman, 1998; Byrd, 2000; Cho, 2006; Fougeron & Keating, 1997; and others). In general, the strengthening effects (longer, greater, slower, less overlapped movements) have been found for both pre- and post-boundary locations, with more robust results closer to the boundary and possibly complementary effects further away (e.g. Krivokapić, 2007).

For example, Byrd (2000) looked at /…C_1V_1(#)C_2V_2…/ sequences with both consonants being /m/s, V_1 a schwa, and V_2 an /i/. While she found a robust effect of the strength of the prosodic boundary on the pre-boundary syllable and measures assessing cross-boundary coordinations, only one out of three speakers produced a similar effect on the CV timing in the post-boundary C_2V_2 syllable. For this single subject, stronger boundary meant later

V_2 gesture; in other words, greater interval between the time of peak velocity in the tongue body raising toward V_2 ([i]) and the onset of C_2 gesture marked by zero velocity crossing for lip aperture. Interestingly, Byrd reported that “[i] peak velocity occurred slightly after the [m] closure for the No-Boundary condition and occurred even later for successively stronger boundaries“ (Byrd, 2000: 12). Informally, with stronger prosodic boundaries, V_2 was ‘moving right‘ with respect to C_2.

Cho (2002, 2006) explored the effects of accent, boundary strength and position on a broad spectrum of articulatory parameters in /b_1V_1(#)b_2V_2/ sequences with Vs corresponding to /a/ and /i/. In the data most relevant to our paper (lip closing in #b_2V_2) the effect of boundary strength was robust in all measures: stronger boundary was realized with longer, larger, slower movements and smaller overlap/truncation between the b_2- and V_2-gestures. Cho tested the predictions of a mass-spring gestural model and parameter scaling within this model against his data and concluded that no single dynamic parameter accounts for the observed variability. Although stiffness scaling accounted for the variation most closely, there were changes in displacement not fitting this explanation. As noted by Cho, displacement of the lip gestures might have been affected by V-gestures and thus it was not strictly context independent.

Locality of the effects has also been investigated. The general pattern observed in several studies was that the movements closest to the boundary were affected the most and this effect waned with the distance from the boundary. For example, Byrd et al. (2006), examining the closing and opening tongue tip movements in nVd_1Vd_2V# and #d_3Vd_4VnV sequences with fixed vowels, found that d_2-closing (pre-boundary, removed from the boundary by a V) was longer/slower for one out of 4 subjects. In the post-boundary case, the d_3-opening, forming the V removed from the boundary by one C, showed longer duration and time-to-peak-velocity for 2 out of 4 subjects. This is a less robust pattern than the one for post-boundary d_3-closing movement immediately adjacent to the boundary and produced longer/slower realizations compared to the control for all subjects. Interestingly, compensatory effects were observed: after strengthening of post-boundary d_3 movements, all 4 subjects showed shortening in duration and time-to-peak-velocity for the d_4 closing movement.

1.2. Articulatory patterns in data varying context and low-level prosody

Relatively limited data and analyses of articulatory timing exist within paradigms eliciting more continuous (and less functionally loaded) variation in prosody. Nittrouer, Munhall, Kelso, Tuller, and Harris (1988) analyzed the effect of stress, syllabic affiliation and speech rate on the timing between the jaw V-to-V cycle and upper lip C-movements (closure onsets). Binary variation in stress and speaking rate were elicited in /…aC#a…/ and /…a#Ca…/ sequences (‘#’ corresponds to a word boundary), C being bilabials /p/ and /m/. Coordination in VCV sequences – analyzed as an angle on a phase plane – varied with rate and stress: faster speech was produced with smaller angle, i.e. earlier lip-closure relative to the V-to-V cycle. Nittrouer (1991) showed similar results when the C-gesture stood for tongue tip raising in alveolar stops. Hence, if C-initiation varies with the jaw-cycle for vowels, even functionally less loaded prosodic variability, such as one related to speech rate, affects the timing in VCV sequences.

Hyper- and hypo-articulation is another non-linguistic dimension with low functional load whose effects on the realization of prosodic boundaries have been investigated (e.g. Cho, Lee, & Kim, 2011; Smiljanic & Bradlow, 2009). Cho et al. (2011) varied the prosodic boundary strength and employed two levels of communicative speech clarity (simulated as speaking to a native vs. non-native speaker) in a#pV Korean sequences. The acoustic analysis found a robust effect of “clear” speech in terms of slower rate and greater number of prosodic units, and also local lengthening of pre-boundary /a/ and post-boundary syllables. Hence, articulatory strategies for slow and hyper-articulated speech were similar and overlapped to some extent with those used for local prosodic boundary marking. There was also an interaction between the prosodic and communicative factors: slowing associated with clear speech was realized more robustly in the vicinity of strong (IP) boundaries. Cho et al. (2011: 357) also reported that “despite the remarkable similarities arising with communicatively driven vs. prosodically driven hyper-articulation, differences between them suggest that the different sources of hyper-articulation are indeed differentially manifested in phonetic output, supporting the view that they are phonetically encoded separately in speech planning”. Hence, this acoustic study suggests a non-trivial relationship between high-level discrete prosodic structure and low-level discrete articulatory precision worth exploring with more continuous articulatory data.

Inter-gestural timing is affected not only by prosodic but also by articulatory context. Löfqvist and Gracco (1999) recorded non-sense VCV sequences including *iba*, *ipa*, *abi*, and *api* in a constant carrier phrase not intentionally varying prosody. The sequences were not immediately preceded by a consonant. The key observation of this study for the current paper was that the lip closure movement started relatively earlier with respect to the tongue body movement towards V_2 for *abi*/*api* than for *iba*/*ipa* sequences.

Šimko, O'Dell, and Vainio (this issue) investigated a dependency of intergestural timing of pV_2 coordination in (Finnish) C_1V_1(p)pV_2 sequences with an alveolar /t/ and bilabial /p/ as C_1 and vowels /a/ and /i/ on the place of articulation of stop consonant C_1. When C_1 was /t/, not directly interfering with the lip movement for /p/, the lip movement started earlier relative to the transition towards vowel V_2 than when C_1 was another bilabial. In sequences starting with /t/, the dependence of pV_2 timing on the vocalic context corresponded to the finding of Löfqvist and Gracco (1999) reported above. However, in sequences starting with /p/ the vocalic context effect was much weaker or actually *reversed* for some speakers. Hence, the timing of consonantal and vocalic gestures in VCV sequences is also affected by the presence of another homorganic consonantal gesture preceding the target consonant.

Similar reversal of CV timing in Slovak *abi*/*iba* sequences preceded by a bilabial has been identified by Šimko, Cummins, and Beňuš (2011) in their analysis of the interplay between vocalic context and speech rate variation. The authors analyzed this inter-gestural coordination in a corpus containing only two sentences but with multiple repetitions under continuous variation of tempo and articulatory precision (largely overlapping with the ones used in the present work). One of their conclusions was that the difference in relative timing of lips and the tongue body movements in *abi*/*iba* sequences is not (entirely) due to a discrete phonological difference between the two sequences but that an important role is played by contextually elicited synergies among articulators. The data used in their analysis support the existence of a lawful, efficient relationship between the timing of gestures, *contextually determined* states of articulators at the onset of these gestures and forces applied to realize the movement.

The studies discussed above looked at the interplay between prosodic structure, segmental context, and inter-gestural timing in somewhat restricted sense, primarily due to the limitations of their datasets. In an effort to provide a fuller picture of articulatory signatures of prosody, we explore the relationship between continuous tempo variation and emergence of prosodic breaks and how these prosodic characteristics relate to the contextual effect of intergestural timing of both consonantal and vocalic gestures in the vicinity of prosodic breaks.

1.3. Modeling articulation at prosodic boundaries

Formal modeling plays an important role in efforts to understand the relationship between prosodic structure and articulation. With the continued progress in understanding the temporal and dynamic nature of speech production (e.g. Browman & Goldstein, 1995), it is natural to explore how dynamically-based modeling responds to the challenges represented by the contextual and prosodic requirements reviewed in the previous two sub-sections.

One of the prominent efforts in this area is the π-gesture model (Byrd & Saltzman, 2003; Saltzman, Nam, Krivokapić, & Goldstein, 2008). The original idea of boundary adjacent adjustments as local stiffness decrease (Byrd, Kaun, Narayanan, & Saltzman, 2000) was refined by Byrd and Saltzman (2003). They model phrase boundaries as π-gestures modulating temporal fabric of an utterance by dynamically adjusting "clock-rate" in their vicinity. Simulations involving slowing down the activation functions of articulatory gestures predicted the observed phrasal effects of longer duration, greater magnitudes and decreased inter-gestural overlap. The model can generate continuous, discrete, and asymmetrical patterns corresponding to the reported articulatory signatures of phrasal boundaries. However, data that serve as a basis for the evaluation, and support the predictions of the model, came from studies employing discrete elicitation paradigms.

The approach to the relationship between continuous aspects of articulatory patterns and more granular/discrete aspects of prosodic structure, e.g. in terms of boundary strength, can be seen as that of separation in this model. Π-gestures as articulatory signatures of prosodic structure are taken to be separate from other articulatory units, which allows abstracting the prosodic structure (phonology) from its realization (articulation).

Two other dynamically based models, despite not being specifically designed or tested with prosodic data, provide different avenues for understanding this relationship. The first one (Gafos, 2006; Gafos & Beňuš, 2006) construes our knowledge of speech as a cognitive system, in which continuous phonetic parameters interact with discrete phonological parameters within non-linear dynamics. The landscape of such interaction describes stable regions as attractors representing more granular phonological aspects, their stability against minor fluctuations of phonetic control parameters, and non-linear bifurcations between these regions when control parameters cross critical values. For articulatory signatures of prosodic structure, one would, within this model, search for a control parameter, representing a continuous phonetic scale, e.g. phasing between gestures crossing a prosodic boundary, and a non-linear equation such that the stable solutions represent the desired number of discrete boundary strengths. The availability of continuous data provides an opportunity for initial exploring of discontinuities and suitable control parameters for modeling the relationship between prosodic structure and articulation.

The problem of relating the prosodic structure and its realization can be approached differently in the optimization-based embodied task dynamics (ETD) platform (Šimko & Cummins, 2010, 2011). In this model, the details of intergestural timing are not specified externally in the form of, for example, phonological rules or non-linear dynamics, but arise from interplays between competing demands on production efficiency and perception efficacy (see Section 5). The model identifies parameters underlying the emergent patterns of relative phasing among gestures (broadly equivalent to the parameters of π-gestures) that capture both the context-dependency and variations in speaking rate and clarity. Prosodic characteristics can be accounted for within this approach through localized changes in relative demands on efficient production, articulatory precision and temporal cohesion among the sequenced gestures.

It is important to point out, that the ETD platform is not conceptualized as a model of online speech production as, for example the task dynamical implementation of Articulatory Phonology (Browman & Goldstein, 1992) and the π-gesture, model. Rather, it is intended to provide a platform for investigating qualitative aspects of inter-gestural coordination that arise from competing requirements of production efficiency, perceptual efficacy and temporal cohesion. Inspired by Lindblom's H&H theory (Lindblom, 1990) and his program of Emergent Phonology (Lindblom, 1999), the predictions of ETD are best interpreted as potential coordination patterns that would arise during phylo- and onto-genetic processes given the validity of the underlying assumption that speech is an efficient rendering of communicative intentions satisfying the constraints of human embodied production and perception apparatuses.

Modeling effort in the current paper focuses on testing the correspondence between the observed data and the ETD model regarding the relationship between continuous measures of speech rate and hypo–hyper articulation and contextual effect of vowels and adjacent labials on the inter-gestural timing at the vicinity of prosodic boundaries. Although not tested, the notions from the other models, such as local slowing down (π-gestures) or the search for suitable continuous parameters and discontinuities pointing to the discreteness in the prosodic structure, also motivate our current effort and will be discussed.

1.4. Motivation summary

The brief review presented above shows that the effects of prosodic structure on articulation have mostly been examined in CV♯CV sequences, eliciting data of discrete prosodic variation, typically induced by varying discrete prosodic structure through syntactic or pragmatic contrasts and controlling for the segmental context. We take a different approach and examine the relationship between prosodic structure and articulation in VC(♯)VCV sequences, in which unintentional emergence of prosodic boundaries is induced by continuous variation in the dimensions of tempo and hypo–hyper articulation. In the absence of discrete intentional variation, the presence of a glottal closing gesture is used for assessing the strength of prosodic boundaries independently of the supraglottal articulatory patterns. We introduce a full control of the rhythmic structure, a partial control for the syntax/semantic structure of the stimuli, and vary the vocalic context. We believe that this kind of data can be informative in search for better understanding of the dynamic parameters utilized by prosody as the phonetic manifestation of the prosodic structure through these low-level variations can be expected to span a wide range of boundary strengths.

More specifically, we examine how phasing relations relate to the contextual influences from asymmetric vowels and how they respond to the temporal distance between homorganic consonants in utterances where both of these contextual effects are affected by prosodic variation. We hypothesize that increasing boundary strength should weaken the influence of the surrounding gestures and phasing patterns should reflect it. Hence, prosodic structure is hypothesized to provide an “affordance” for different (either quantitatively or qualitatively) efficient and synergistic realizations of temporal sequencing patterns.

Recall the relationship between relative timing of gestures in CV sequences with a bilabial consonant reported in Section 1.2. The phasing details were shown to depend on the articulatory context, namely on whether the preceding consonants shared the same articulators with the bilabial. In this work we investigate whether prosodic structure provides a continuous counterpart of this binary effect. Presence or absence of an interfering articulation is replaced by a continuously decreasing or increasing temporal interval separating two bilabials. The relative duration of the interval is hypothesized to be associated with the strength of prosodic boundary.

The paper also compares the trends identified on continuous intervals of prosodic variation with the patterns of the relationship between boundary strength and articulation observed in data elicited with the discrete paradigm. Finally, we test how these trends correspond to the predictions of the ETD model and thus to what extent they can be interpreted as emergent from interplays between articulation, perception and communicative intentions.

2. Materials and methods

2.1. Subjects

Four native speakers of Slovak, one female and three males, participated in this study. Slovak was chosen because target asymmetric VbV sequences *abi* and *iba* represent real and very frequent Slovak words: *iba* means ‘only’ and *abi* means ‘so that’.

2.2. Material

In order to assess and extend the findings reported in Section 1.2 and include contextual variation in target sequences, two asymmetric (♯)VCV sequences were chosen: (♯)*abi* and (♯)*iba*. Each target word was embedded in a semi-meaningful sentence to achieve naturalness of the stimuli and as close as possible similarity between the two sentences in terms of syntactic and prosodic structures. In order to enable observations of the vocalic lingual movements and to minimize the effect of surrounding consonants, the target sequences were flanked by bilabial nasals /m/, preceded and followed by vowels with contrasting canonical tongue requirements. Hence, the two target sequence were: /… im(#)**abi**mu…/ and /… am(#)**iba**mu…/. The two stimuli sentences with IPA transcriptions and transliterations are shown in (1) and (2).

Cítim *aby* *mu* *krásne* *pristali*.
[t͡siːcim ɑbi mu krɑːsɲɛ pristali]
‘Feel−1stSg so that him nicely fit−3rdPl’ (1)

Čítam *iba* *mu* *krásne* *po grécky*.
[t͡ʃiːtɑm ibɑ mu kraːsɲɛ pogrɛːt͡ski]
‘Read−1stSg only him nicely in Greek. (2)

Note that, [ɑbi] serves here as a conjunction and thus requires a sentence with two clauses while [ibɑ] is typically analyzed as a clausal particle and requires a single clause. The two stimuli thus do not achieve the complete syntactic identity. Although the correspondence between syntactic and prosodic boundary marking is never perfect and may be influenced by semantic, discourse and phrase length aspects (e.g. Bachenko & Fitzpatrick, 1990), syntax provides a strong structural influence on the distribution of prosodic breaks, especially in read speech (e.g. Nespor & Vogel, 1986; Selkirk, 1986; Truckenbrodt, 1999). Hence, the syntactic structure favors a stronger prosodic break before *abi* than *iba*. Moreover, there is a stronger tendency for glottalization to occur before low than non-low vowels, presumably linked to preferred articulatory setting and perceptual consequences of glottal closure followed by low vowels (Brunner & Zygis, 2011). Syntactic and articulatory-perceptual factors thus predict greater presence of glottalization before [ɑbi] in (1) than before [ibɑ] in (2). However, in both sentences, a prosodic break before the target VCV word is felicitous and natural in Slovak. On the contrary, a prosodic break after the target word is not possible and did not occur in our data, especially as the third word *mu* is an enclitic prosodically aligned with the preceding word.

2.3. Elicitation of continuous variation in tempo and precision

For each block of data collection, subjects were asked to read repeatedly the stimuli sentence shown on a screen in regular Slovak orthography. The method used to elicit variation in speaking rate and speech clarity was inspired by the paradigm introduced by Cummins (1999). In one block,

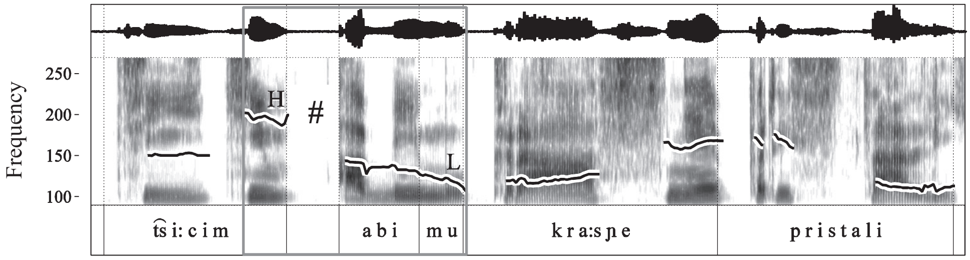

Fig. 1. A sample prompt token in (1) with the oscillogram, the spectrogram with overlaid F0, and the alignment of words to the signal. 'H' and 'L' mark F0 high and low targets and '#' marks a prosodic boundary realized as a silent interval in this token.

speech rate was varied by an experimenter signaling a desired rate by moving an indicator along the axis between 'extremely slow' and 'extremely fast'. The position of the indicator was changed randomly after every 3–4 repetitions encouraging the speaker to probe the entire spectrum of potential rate variation. In the second block, the same was repeated with the axis of 'extremely precise articulation' and 'extremely relaxed articulation' thus encouraging the subject to range from very lax, hypo-articulated productions to very clear, hyper-articulated ones. Subjects were instructed to speak fluently, not divide the sentence into multiple chunks, and to keep the speech rate of the sentence in the second block as stable as possible. More specifically, the speakers were asked to maintain a "natural" speaking rate and a "natural" clarity during precision and tempo variation conditions, respectively.[2] These procedures ensured that the resulting data contained substantive variation in both rate and articulatory precision.

2.4. Data collection

Both articulatory and acoustic data were collected using the same procedure as described in detail in Beňuš (2012). Briefly, electro-magnetic articulography (EMA, AG500, Carstens Medizinelektronik, IPS Munich) was used for tracking the movements of seven coils attached to active articulators in a mid-sagittal plane: the upper lip, lower lip, the lower incisors to record jaw movement, and four sensors on the tongue (TT, TB1, TB2, and TD) glued in roughly equidistant intervals between the tongue tip area and the velar/dorsal region of the tongue. In addition, four reference sensors were attached and used to correct for head movement: two sensors behind each ear, one on the nose, one above the upper incisors (Hoole & Zierdt, 2010). Movement data were filtered with 60 Hz for the tongue tip sensor, 20 Hz for all other sensors attached to the active articulators, and 5 Hz for the reference sensors. Acoustic signal was captured with a directional Sennheiser MKH 40 microphone with a sampling rate of 32,768 Hz and downsampled during post-processing to 16,384 Hz.

2.5. Data labeling

Fig. 1 shows one rendition of the prompt sentence (1). The gray rectangle marks the target sequence /… im(#)abimu…/ that is underlyingly voiced.

Note that the target sequences are underlyingly voiced (im(#)abimu/am(#)ibamu). For the present purposes, any deviation from normal modal voice was taken to signal a disjuncture between the initial verb and the following target word and increase the saliency of the prosodic break.[3] Glottalization accompanying vowel-initial words is known to correlate with the strength of the preceding prosodic boundary and with the degree of prosodic prominence on these words (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996; Pierrehumbert & Talkin, 1992). In our corpus, pitch accents on the target *iba*/*abi* words are extremely rare. Moreover, the interval for break affordance was always preceded by a high F0 target (Fig. 1). Since F0 did not fall into the low regions of the speakers' pitch ranges, glottalizations cannot be associated with lowering of F0 marking the boundary (e.g. Pierrehumbert & Frisch, 1994). Hence, glottal activity in this region is attributable neither to pre-boundary F0 lowering nor to the presence of pitch accents on the V-initial target words. Deviations from vocal cords periodicity were taken to complement (H)igh boundary tones (De Pijper & Sanderman, 1994).

As we do not have access to articulatory data depicting the glottal movement, we rely on the acoustic signal. The tokens were labeled by an experienced labeler for modal voice (MV), glottalization (Glot) or Silence (Sil), and the duration of the non-modal interval for Glot and Sil tokens was recorded. Additionally, data presentation and results include a naïve measure of speech rate as the interval between the acoustic release of target /b/ and the end of the utterance.

Articulatory signal was labeled in a semi-automatic procedure illustrated in Fig. 2. In this work we focus on two articulatory movements in our target sequences /m(#)abi/ and /m(#)iba/ that are substantially co-produced: the lip opening and closing gestures between /m/ and /b/ and the tongue body transition from the target of the first vowel to the target of the second vowel. In order to identify salient kinematic characteristics of these movements, a lip aperture measure (referred to as LA) was calculated as the Euclidean distance between the positions of the lower lip and upper lip sensors. The onsets and offsets of the lip movement (*M-offs*, *B-ons*, *B-offs*) were identified as velocity zero-crossings of the lip aperture signal (using zero-crossing instead of thresholds related to peak velocity was facilitated by the segmental structure of our stimuli). Similarly, the onsets of the tongue body movement (*V2-ons*) were placed at the velocity zero-crossings of the first principal component of the TB2 signal and correspond to the times of maximal displacement of the tongue body approximating a high target for the vowel /i/ and a low target for /a/ in an /iba/ sequence. Times of peak velocity achieved by the articulators were also recorded (*M-pv*, *V2-pv*).

In total, 952 tokens were labeled for both acoustic and articulatory information, their distribution among the subjects and target VCV sequence is shown in Table 1.

2.6. Data processing

To facilitate the discussion of dynamic characteristics of the temporal organization of articulatory movements and the effect of prosody, we employ a phase of gestural movement as the main tool of assessing temporal inter-gestural coordination (cf. Kelso & Tuller, 1985).

[2] As it is common to hypo-articulate in faster speech, the complete separation of tempo and hypo–hyper dimensions was not achieved, which is discussed in Section 3.1.

[3] A type of creaky voice sometimes called vocal fry is possible for very high F0 targets, which, however, was not the case in our data.

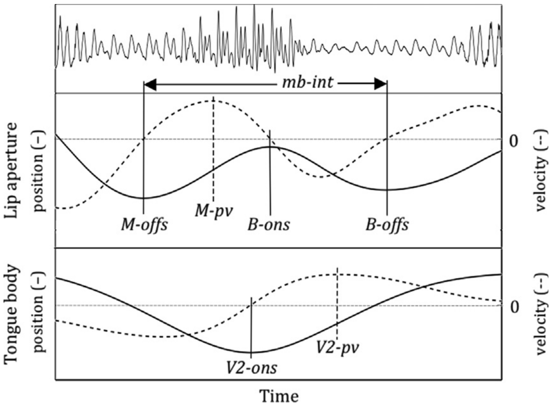

Fig. 2. Articulatory labeling for an *abi* sequence (approximately /im(#abi)/ is shown). Audio signal, positions (solid curves), and velocities (dashed curves) of LA and TB2 are shown. Relevant onsets and offsets of gestures were identified as velocity zero-crossings (0 velocity axes are plotted for reference). Instants of peak velocity for lip opening and lingual gestures were also identified.

Table 1
Distribution of two target sequences among the subjects.

Subject	abi	iba	Total
S1	136	121	275
S2	146	145	291
S3	132	127	259
S4	50	95	145
Total	464	488	952

Temporal characteristics of the movement of a mass under the influence of a particular (second-order) dynamics are contingent on parameters of the underlying dynamics. Based on their values, the landmark states, such as the moment when the mass moves with the highest velocity, are achieved at various times, but at the same phase of the movement's duration.

For undamped second order mass-spring dynamics, for example, the phase angles at which the (positive and negative) peak velocity is achieved are $\pi/2$ and $3\pi/2$. Undamped oscillators, however, are not a suitable tool for modeling most articulatory movements that are target oriented and are more precisely approximated by damped oscillatory trajectories. A critically damped mass-spring dynamical system approaches its given dynamical target in the speediest possible way without overshooting it (Saltzman & Munhall, 1989). Unfortunately, it is difficult to generalize beneficial properties of phase measurement for critically damped systems. One way of avoiding this obstacle is to characterize the instantaneous state of a damped dynamical system through the phase of its undamped version.

It can be shown, that regardless of the parameter values, the critically damped oscillator reaches its peak velocity at the phase of 1 rad of its undamped counterpart. The peak velocity of the *critically damped* movement thus provides an invariant reference point to the phase of the undamped oscillator with the same parameters. If we define the "phase"[4] of the critically damped oscillator as the phase of its undamped counterpart, the duration of interval from the onset of the movement (its zero-velocity) to the time of peak velocity – *time-to-peak-velocity* – provides a unit (literally, value 1) for its measurement. If *ons* is the time of the movement onset, *pv* is time of its peak velocity and *event* is time at which an event of interest occurs, we can estimate the phase of the gesture *G* underlying the movement at the occurrence of the event as

$$\phi_G(event) = (event - ons)/(pv - ons) \tag{3}$$

We shall use the above definition to estimate the phase values of the event of interest (onset of bilabial closure with respect to surrounding gestures). This measure provides, to an extent, a measure of inter-gestural coordination normalized for speaking rate differences as well as differences in gestural dynamical parameters, e.g., stiffness and dynamical target.

The suitability of the measure depends on several assumptions. First, that articulatory movement is well approximated by critically damped second order dynamics (or, more precisely, a weaker assumption, that the underlying dynamics satisfies the properties used in the definition: a fixed phase of the instant of peak velocity of damped dynamics within an undamped counterpart, and a linear nature relationship between the phase and time). Second, that the time-to-peak-velocity indeed provides a reliable characteristics of the dynamics of the movement under consideration, i.e., that the peak velocity is achieved by purely dynamic means (we shall come back to this point later in the text). As we cannot guarantee that these conditions are always satisfied, whenever felt appropriate we refer to our measurements as *observed phase*.

We will primarily use the following derived measures in our articulatory analysis:

$$\phi_{V2}(bons) = (B{-}ons - V2{-}ons)/(V2{-}pv - V2{-}ons), \tag{4a}$$

$$\phi_m(bons) = (B{-}ons - M{-}offs)/(M{-}pv - M{-}offs), \tag{4b}$$

$$mb{-}int = B{-}offs - M{-}offs, \tag{4c}$$

$$mb{-}int{-}norm = m(\#)b{-}int/part{-}dur. \tag{4d}$$

[4] From now onwards, we shall drop the quotation marks and refer simply to a *phase* (or *observed phase*, see below) of a gesture.

In these equations, *B-ons* and *V2-ons* are articulatory onsets of gestures /b/ and V_2, respectively, *B-offs* and *M-offs* are the offsets of closing bilabial gestures after completion of the closures, *V2-pv* and *M-pv* are times of peak velocity of gestures V_2 and the opening gesture following /m/ (see also Fig. 2), and *part-dur* is the duration of the interval from the end of (acoustical) closure for /b/ to the end of the utterance. Hence, $\phi_{V2}(bons)$ in (4a) represents the onset of b-closure as the phase of V_2 movement. $\Phi_m(bons)$ in (4b) corresponds to the onset of b-closure as the phase of m-opening movement. *mb-int* in (4c) is defined by the articulatory offset of /m/, i.e. the onset of *m*-opening, and the offset of /b/, i.e. the end of b-closing. However, raw *mb-int* includes both temporal aspects of global speech rate variation as well as local slowing assumed to be controlled by the prosodic boundary. To estimate the local temporal aspects, (4d) normalizes *mb-int* for the duration of the prompt sentence after /b/. We took the longest possible continuous interval excluding m(♯)b-interval for normalization rather than using the entire duration of the prompt sentence to avoid the part-whole problem (e.g. Benoit, 1986). The variable *mb-int-norm* thus captures tempo-normalized duration of m(♯)b-interval and can be used as a measure of localized slowing down (or speeding up) relative to the rest of the utterance in the vicinity of structural affordance for a prosodic break. Moreover, the modeling effort presented in Section 5 is based on localized temporal changes. Hence, analyzing the relationship of prosodic structure and temporal sequencing of gestures through the normalized time allows for a more direct comparison between the modeled and observed patterns.

2.7. *Data presentation*

One of the goals of our analysis is the examination of trends in the relationship between prosody and articulatory/dynamic organization (Section 1.4). For this reason, we primarily concentrate on exploring these trends in the data pooled for all subjects. Additionally, pooling subjects together gives a better and more complete picture of the prosodic continuum under investigation since we include variability both within and across the subjects. Nevertheless, most of the results we present are robust also separately for speakers.

When continuous data are presented, our focus lies in examining the directions of the trends (fits) and not so much in assessing their quality and/or power. This is because data collected in this less controlled paradigm are highly noisy and our goal is to understand the continuous nature of the patterns. Additionally, we are interested in extrapolation between the current data, sampled in a biased way from weaker prosodic boundaries, and patterns in the literature and model predictions that concern primarily stronger prosodic breaks.

3. Results

3.1. *Variability in tempo and precision*

We first describe the achieved variability along the two dimensions targeted by the elicitation procedure: tempo and articulatory precision. Fig. 3 shows separate plots for the two target words (*abi*, *iba*) and four subjects using a raw measure of tempo as the duration of the utterance and a raw measure of precision as *HH-index*, which is a sum of standard deviations of vertical and horizontal positions of 7 sensors attached to the active articulators during each utterance.

Several observations can be made. First, the procedure was relatively successful in eliciting continuous variation along both dimensions. For most speakers there is a considerable difference in the measure relevant for the given condition, i.e. overall utterance duration and *HH-index* for tempo and precision variation, respectively. The largest recorded values of the measure are approximately twice the smallest ones. Second, the large overlaps between the two ellipses, the widths along their minor axes, as well as tilting of the precision (gray) ellipses away from the horizontal plane shows that subjects were not able to completely separate the two dimensions. Hence, considerable variation of tempo was produced during varying precision, and relatively smaller variation of precision was produced during varying tempo. This is in line with results of Smiljanic and Bradlow (2009: 245–246) reporting intertwined effects of temporal variation and speech clarity as well as considerable inter-subject differences in strategies for clear speech reported e.g. in Perkell et al. (2002).

Given the different elicitation procedures and subsequent differences between the distributions, we report most of the observations separately for tempo and precision blocks. As we shall see, despite these differences, the relationship between prosodic boundary strength and temporal patterning of gestures in our data is qualitatively similar in the two data sets.

3.2. *Presence vs. absence of a break*

Our corpus provides continuous prosodic variation and the production of utterances with internal prosodic breaks was neither sought nor intended. Rather, subjects were instructed to speak fluently. Despite this, they signaled a disjuncture between the sentence-initial verb and the target *iba*/*abi* with non-modal activity of the vocal folds (either present as glottalization, whisper, or complete silence in the acoustic signal, or combination of these) relatively frequently (37% of all the tokens, see Table 2).

The distribution of prosodic breaks in the data is summarized in Table 2 that divides breaks by block (tempo vs. precision), target string, and whether the break was realized only as (partial) glottalization or whether it included a silent interval as well. Pooling all subjects together, breaks were significantly more frequent before *iba* than *abi*, which was more robust in the tempo variation than in the precision variation; $X^2[1]=32.95$, $p<0.001$ and $X^2[1]=5.48$, $p=0.019$.[5] Finally, while this pattern applies to both glottalizations and silent intervals for tempo variation, in precision, *abi* tokens tended to be realized with silences and *iba* tokens with glottalizations.

3.3. *Breaks and time are intimately linked*

Our annotations provide two measures of boundary strength as a key parameter representing prosodic structure: localized slow down assessed with supraglottal articulatory movements (*mb-int-norm*) and discrete settings for vocal cords activity (MV, Glot, Sil). Fig. 4 shows the relationship between them separately for target string and blocks varying tempo and articulatory precision.

[5] The reported significance of the difference between *abi* and *iba* in the pooled data should be taken with some caution in the view of subject variation. It should be noted that in S2 data, showing the most robust pattern of more frequent breaks before *iba* than *abi*, breaks produced as glottalizations were the major source of this difference, and S4 showed more frequent breaks before *abi* than *iba*.

Please cite this article as: Beňuš, Š., & Šimko, J. Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* (2014), http://dx.doi.org/10.1016/j.wocn.2013.12.005

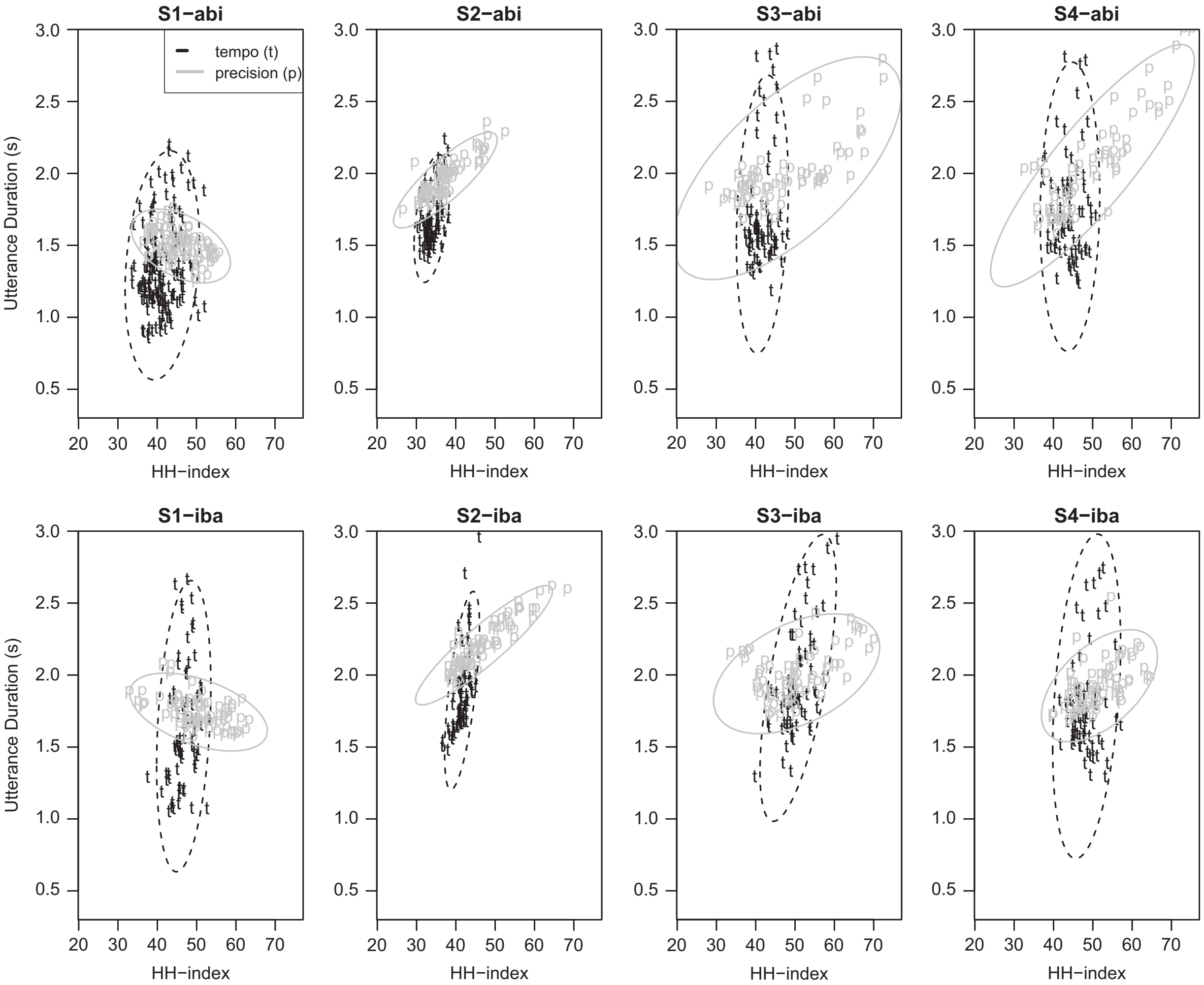

Fig. 3. Variability along articulatory precision (*x*-axis) and tempo (*y*-axis) separately for subjects and prompt sentences. Ellipses show 95% confidence interval; entire data set ($N=1210$ tokens).

Table 2
Distribution of prosodic breaks (no-break vs. break) by type, target string, and whether the break was realized only as glottalization (Glot) or also included silence (Sil).

String	Tempo								Precision							
	No-break		Break				Total		No-break		Break				Total	
			Glot		Sil						Glot		Sil			
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
abi	193	39.3	33	6.7	16	3.3	242	49.3	141	30.6	14	3	67	14.5	222	48.2
iba	137	27.9	66	13.4	46	9.4	249	50.7	125	27.1	62	13.4	52	11.3	239	51.8
Total	330	67.2	99	20.1	62	12.7	491	100	266	57.7	76	16.4	119	25.8	461	100

All four subplots show the longest normalized m(♯)b-intervals for tokens with the strongest breaks (silence) and the shortest ones for the weakest/non-break MV tokens. A mixed-models test with subject treated as random factors showed a robust main effects of both vocal cord activity ($F=320$, $p<0.001$) and target string, $F=339$, $p<0.001$) showing greater local slow-down in *abi* than *iba*. The interaction between these two factors as well as the effect of block (tempo vs. precision) were much weaker ($F=13.9$, $p<0.1$ and $F=12.8$, ns. respectively).[6] Our data thus exhibit the expected positive relationship between locally adding time (*relative slowing*) and boundary strength (vocal cord activity) and show that *mb-int-norm* is thus a suitable measure of boundary strength. Despite robust differences among the three categories, the boxplots show massive overlaps in that the majority of tokens could fall into more than one category.

Fig. 5 illustrates the relationship between glottal and supraglottal gestures in a slightly different way by plotting *mb-int-norm* and raw duration of the non-modal voice interval (for Glot and Sil tokens). Since the non-modal voice interval is in fact contained in the m(♯)b-interval, the strong positive correlation was expected. Importantly, the plots illustrate continuity both within and across the two categories. Moreover, the precision block provides less noisy data for examining the articulatory signatures of boundary strength induced by low-level tempo/precision variability.

[6] Here and in further mixed-models tests, Monte-Carlo simulations for determining *p*-values in R was used; lme4 package was used for mixed-models testing.

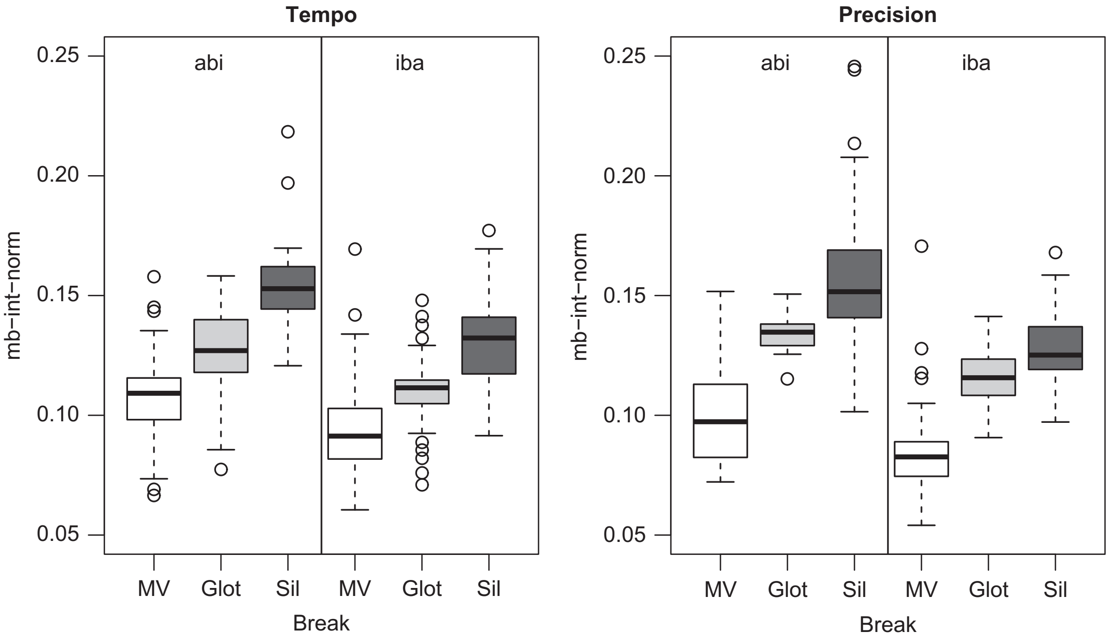

Fig. 4. Normalized duration of the interval between /m/ and /b/ split by the target VCV sequence (*abi*/*iba*) and three-way categorization of vocal cord activity ('MV' modal voice, 'Glot' glottalization, 'Sil' Silence) separately for tempo and precision variation.

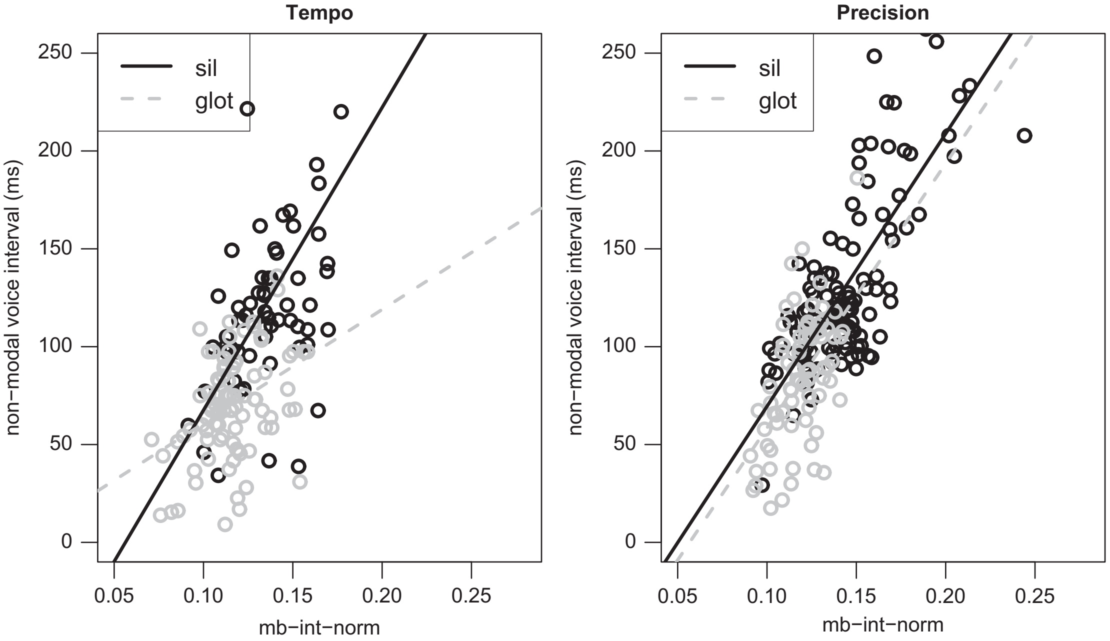

Fig. 5. Duration of the non-modal voice activity for the vocal cords against the normalized m(#)b-interval separately for the tempo and precision blocks and the two non-modal settings of glottalization (Glot) and complete closure (Sil).

3.4. Phasing between /b/ and V_2

Next we examine how *mb-int-norm* (as a measure of boundary strength) relates to temporal patterning and vocalic environment. We use a time normalized measure of the *phase* of V_2-gesture at the instant of the onset of *b*-gesture ($\phi_{V2}(bons)$) introduced in Section 2.6 as a measure of relative timing between the gestures of interest. Fig. 6 presents the data.

Examining the boxplots first, an overall mixed-models test for the effects of Break (MV, Glot, Sil), String (abi, iba), and Block (tempo, precision) with Subject as a random factor shows significant effects only for Block ($F=105.6$, $p<0.01$), String ($F=44.2$, $p<0.01$), and their interaction ($F=24.3$, $p<0.01$). Subsequent separate tests for the effect of Break in each String show a significant difference between MV and Glot ($p<0.001$, $p=0.002$) and between Glot and Sil ($p=0.04$, $p<0.001$) respectively for *abi* and *iba*. Hence, increasing boundary strength is realized with more negative phase both in tempo and precision blocks. Therefore, stronger boundaries led to b-onset leading the V_2-onset more than in weaker boundaries.

The scatter plots in the bottom row show the relationship between our continuous measure of boundary strength, *mb-int-norm*, and bV_2 phasing assessed as $\phi_{V2}(bons)$. We observe that as *mb-int-norm* gets longer (and boundary presumably stronger), the lip-closing for /b/ starts earlier relative to V_2.This pattern is supported with significantly negative regression slopes for three out of four tests: tempo-*abi* ($t=-5.15$, $p<0.001$, $R^2=0.1$), precision-

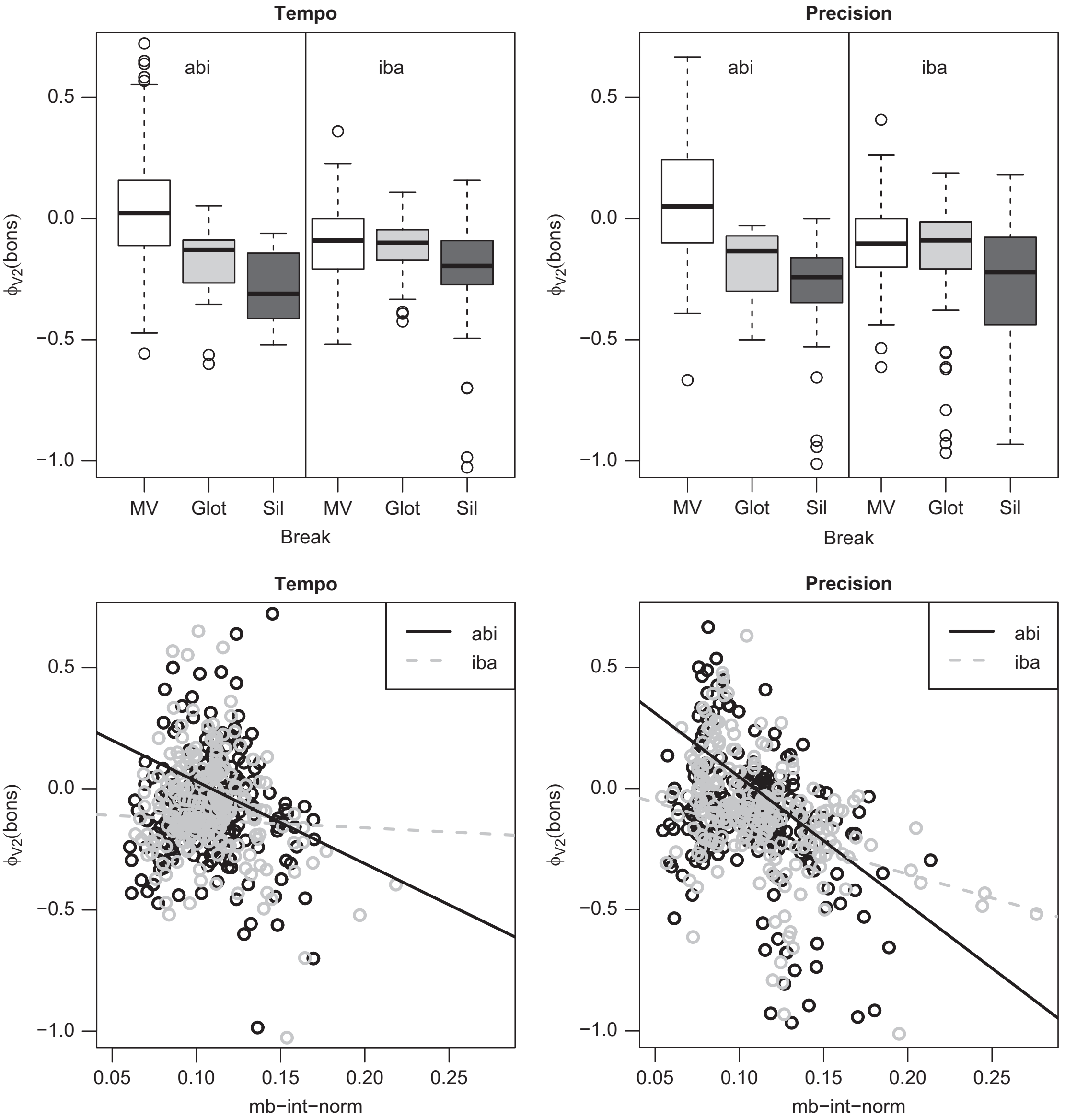

Fig. 6. Relative timing of /b/ and V_2 divided by the modality of voice during the break, target string, and block in the top row boxplots, and b–V_2 phasing as a function of normalized m(♯)b-interval and divided by target string and the block in the bottom row scatterplots.

abi ($t=-14.14$, $p<0.001$, $R^2=0.47$), and precision-*iba* ($t=-3.35$, $p<0.001$, $R^2=0.04$). This suggests a stronger relationship for *abi* than *iba* and for precision than tempo block.

3.5. Phasing between /m/ and /b/

Due to the nature of our procedure and material, adjacent labial consonants /m/ and /b/ are subject to different pressures arising from prosody (tempo, precision) compared to the lip and tongue gestures for /bi/ or /ba/. This is because in the former, these adjacent gestures control the same tract variable and have a strict precedence: the lips first need to open after m-closure and only then can start closing for the b-closure. In the later, the timing of C- and V-gestures is less constrained physiologically: the lips and the tongue gestures only share the jaw.

Phasing between m-opening and b-closing is conceptualized here in terms of ϕ_m(*bons*), i.e., the onset of b-closing as the phase of m-opening gesture (see Section 2.6). The top row in Fig. 7 contains the boxplots of the phase values for the three voice modalities split by the type of elicitation (tempo, precision) and target sequence (*abi*, *iba*). The four plots in the second and third row show the scatteplots for *mb-int-norm* and ϕ_m(*bons*), again separately for target sequences, elicitation type, and vocal cord activity types.

Most of the boxplots indicate non-linear relationships: the phase is the highest (b-closing starts relatively latest) for modal-voice tokens, but for the remaining tokens the median phase is greater (or equal) for those with silence (stronger break) than for those with glottalization. The behavior suggested by the plots adds further detail to this observation: for tokens with no non-modal boundary small values of *mb-int-norm* are associated with high ϕ_m(*bons*) and the (solid black) linear regression lines have negative slope, significant for three out of four such lines ($t=-3.94$, $p<0.001$ for tempo-abi, $t=-9.0$, $p<0.001$ for precision-abi, and $t=-3.53$, $p<0.001$ for precision-iba). For tempo-iba the sign of the slope was not significant ($t=-1.59$, $p=0.11$). Furthermore, at least for the two precision (rightmost) scatterplots, significantly negative slopes for the tokens with modal voice are complemented with significantly *positive* slopes for the tokens with silence (black dotted lines; $t=2.88$, $p<0.01$ for *abi* and $t=1.95$, $p=0.05$ for *iba*).

Please cite this article as: Beňuš, Š., & Šimko, J. Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* (2014), http://dx.doi.org/10.1016/j.wocn.2013.12.005

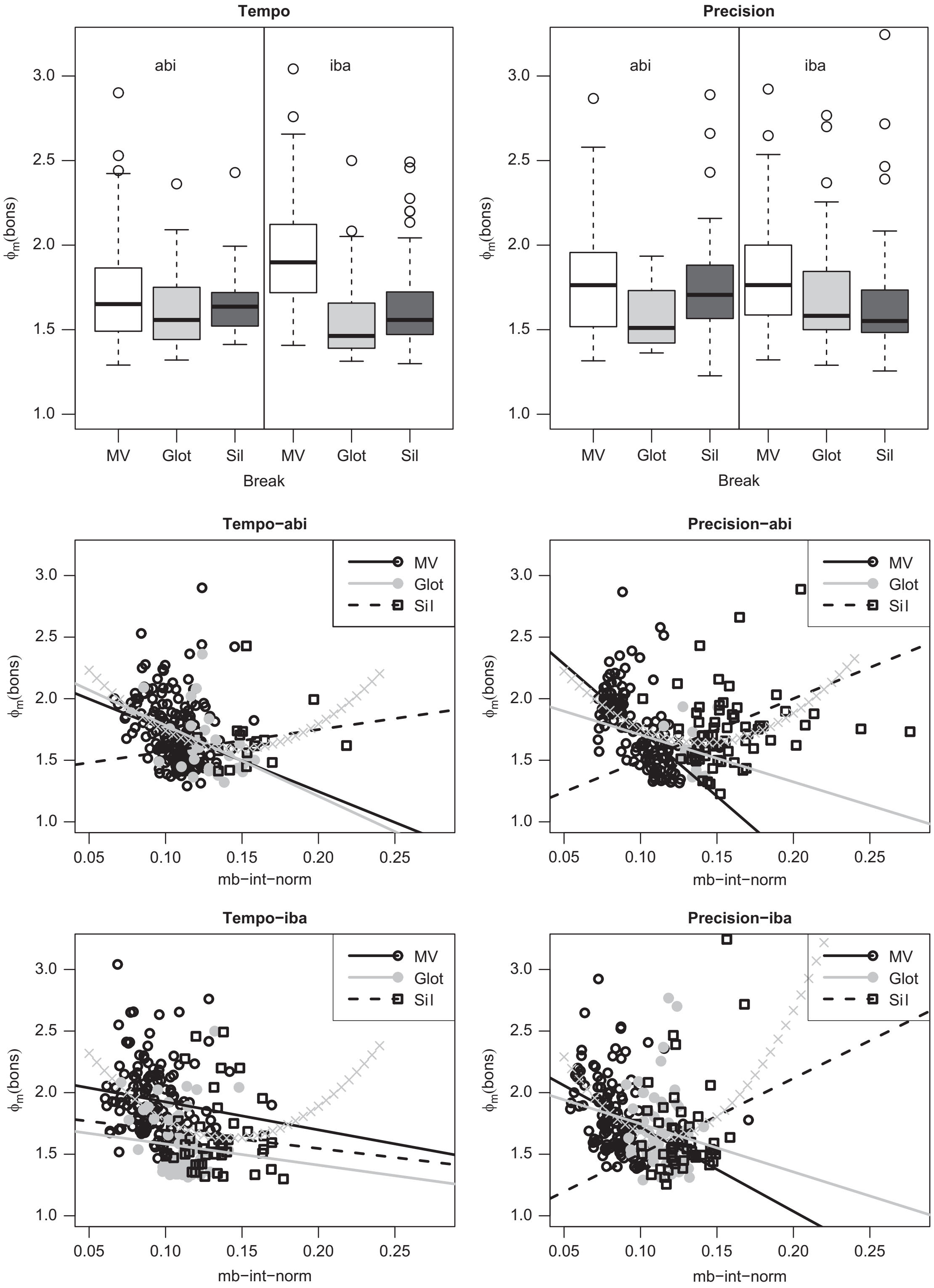

Fig. 7. Relative timing of m-opening and b-closing divided by the modality of voice during the break; modal voice (black solid line), glottalization (gray solid line), silence (black dashed line). The quadratic fits shown with gray '+' marks cover all the data.

This bimodality of the relationship between local time at the prosodic boundary and phasing of the two bilabial gestures across this boundary is supported by the fact that for all four scatterplots, the quadratic fits for the pooled data, shown with gray '+' marks, reached higher R^2 values than corresponding linear fits (clockwise quadratic: 0.11, 0.16, 0.14, and 0.15 and the respective linear: 0.06, 0.0, 0.12, 0. 06). ANOVA tests for all four comparisons show that the quadratic model provides a significantly better fit than the linear one ($F=14.5$, $p<0.001$, $F=45.1$, $p<0.001$, $F=9.2$, $p=0.002$, $F=23.8$, $p<0.001$). Finally, we observe intermediate values of slopes for precision tokens with glottalizations (solid gray lines) that are assumed to correspond to intermediate boundary strengths. Taken together, our data suggest that the relationship between m-b phasing and local time at the syntactic affordance is non-linear.

Please cite this article as: Beňuš, Š., & Šimko, J. Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* (2014), http://dx.doi.org/10.1016/j.wocn.2013.12.005

3.6. Phasing relation and raw vs. normalized m(♯)b-interval

Previous sections presented the relationship between the normalized m(♯)b-interval on the one hand and CV-phasing and CC-phasing on the other. The choice of using the normalized interval in Sections 3.3–3.5 was justified in Section 2.6 and allowed us to tap into the prosodic structure as much as possible and filter out the effects of global tempo changes.

In order to assess the difference between the effects of global (phonetic) and local (linguistic) tempo changes on the target gestural phasings, we performed similar analyses to the ones in Sections 3.3–3.5 also with raw m(♯)b-interval. Because the observed macroscopic patterns largely agree with those presented above, and due to space reasons, we will not present the complete analyses here. The major difference between the two measures was that the reported relationship between phasings and m(♯)b-interval were *more robust* for the raw measure than for the normalized one. For example, the R^2 values for the quadratic fits reported in Section 3.5 reached the values of 0.11, 0.16, 0.14, and 0.15 for the normalized interval and 0.22, 0.48, 0.18, and 0.14 for the raw one. Importantly, the observed similarity of the relationship between temporal phasing and tightening–relaxing of temporal demands in the vicinity of the break irrespective of the underlying source of this demand suggests that local and global temporal variations seem to exert qualitatively similar influence on inter-gestural coordination. We will take this point further in the general discussion.

4. Discussion of empirical results

Data for this study were elicited through continuous variation of tempo and precision. This resulted in a considerable number of tokens with an identifiable prosodic break. The price paid for this methodological approach was a corresponding noise in most measurable kinematic and dynamic characteristics including those used in this work. That consistent and coherent patterns were obtained despite this large variability indicates the robustness of the reported phenomena.

Syntactic and articulatory-perceptual reasons predicted a greater presence of glottalization in *abi* than *iba*. This, however, was not observed in our data given the greater frequency of breaks in *iba* than *abi* sentences (Section 3.2). This shows that prosodic variability in terms of disjunctures in this data set is not primarily determined by syntax or universal articulatory-perceptual considerations, but the demands from tempo/precision variability. Syntax-prosody provides ‘affordance’ (in the sense of Gibson (1979)) for the realization of a prosodic break.

Section 3.3 showed that the relationship between the strength of a boundary – expressed with the vocal cord activity, either discretely as vocal cords settings (modal voice/glottalization/silence) or continuously with the duration of non-modal voice interval – is linked to the time available for the transition between /m/ and /b/. In this sense, local slow down, expressed with normalized m(♯)b-interval, provides an assessment of boundary strength: it shows a clear relationship to the independent marking of boundary strength with non-modal activity of vocal cords, and filters the effect of global rate variability. The elicitation paradigm allows a peek into the “birth” of salient prosodic breaks (those including silence) in response to the requirements, more specifically their relaxation, in terms of speech tempo and precision.

One of the goals of this paper is to extend our understanding of the synergistic CV timing effects with respect to the vocalic context as observed by studies reviewed in Section 1.2, and simulated with ETD model (Šimko & Cummins, 2010, 2011) discussed in Section 1.3. Recall that according to these studies the lip closing movement starts relatively earlier with respect to lingual vocalic movement in *abi* than in *iba* sequences, at least when no prosody demands and no interfering preceding bilabial consonant are at play. In the current context, we should expect smaller $\phi_{V2}(bons)$ – earlier relative onset – for *abi* than for *iba* tokens. According to Fig. 6, however, this was generally the case only for tokens with *mb-int-norm* (*x*-axis) greater than approx. 0.15, where the solid black line (*abi*) lies below the dotted gray line (*iba*). For tokens with smaller *mb-int-norm* (weaker boundaries or modal voice) the pattern is reversed and lip movement tends to start relatively *later* for *abi* than for *iba* sequences. We attribute this effect to greater “gestural crowding”, i.e. progressively stronger effect of the homorganic lip gesture for /m/ on the behavior of the lips at the onset of the b-closing gesture as the boundary weakens.[7]

Our results concerning the relationship between boundary strength and post-boundary bV-phasing agree with the literature reporting on the data elicited within other paradigms. Byrd (2000) and Cho (2006) found that the strength of the prosodic boundary affects coordination around prosodic boundaries making the transitions less overlapped. For the CV coordination, 1 out of 3 subjects in Byrd (2000) and all subjects in Cho (2006) employed this mechanism also in the post-boundary ♯CV coordination. The observed trend in the current work – the greater temporal lead of the b-onset relative to the transition to V_2 the stronger the boundary – is consistent with these findings despite the fact that our data are more continuous, biased toward non-existent or weak prosodic boundaries, and CV-coordination in question is not immediately adjacent to the prosodic break. This observation can be conceptualized through the notion of “crowding” mentioned above: the onset of b-closing in tokens without a break is ‘pushed to the right’ due to crowding from the preceding /m/, and the requirement to open the lips after it. Since the V_2-onset is not under similar pressure, the lip-closing for /b/ starts relatively later in relation to the V_2-onset for tokens without a break. In tokens with a break, on the other hand, prosodic structuring creates ’affordance’ for timing pattern similar to the “non-crowded” case observed in contexts without such a pressure from preceding homorganic consonants.

The analysis of the relationship between phasing of the two labial movements and boundary strength suggests a non-linear pattern: in realizations with breaks the two gestures are more overlapped given less time (behavior consistent with the “crowding hypothesis”), but with no breaks, less time resulted in seemingly smaller overlap. This decreased overlap in fast tokens with no boundary could be interpreted as a ‘push’ of the b-closing gesture to the right – phenomenon that could partly explain its later phasing relative to V_2.

This apparent “bimodality”, however, requires further inspection. Our measure $\phi_m(bons)$ might in fact be deficient in capturing the relationship between phasing and temporal demands for tokens with great overlap between adjacent gestures. The reason is that the phase estimate uses peak velocity as a landmark pertaining to the autonomous dynamics of the gesture, but in some cases the measured peak velocity may rather be related to the truncation by the following homorganic gesture. Recall, that $\phi_m(bons)$ is defined as a ratio between the durations of the interval from the onset of the opening movement to the onset of the closing movement (i.e., a two consecutive velocity zero-crossings of lip aperture variable) and the interval from the onset of the opening movement to the instant of maximal velocity of the movement. The observed peak velocity must come before the second zero-crossing, and thus the value of $\phi_m(bons)$ must always be greater than 1. In fact, the measured value must be considerably greater than 1 given

[7] Similar effect of homorganicity of preceding gestures has been reported by Šimko, O'Dell, and Vainio (this issue) who did not vary prosody but directly compared the phasing relationships in sequences with homorganic and heterorganic consonants.

the necessary time for the lip aperture to decelerate from its maximal velocity to 0. Perhaps the smallest values of $\phi_m(bons)$ in Fig. 7 (around 1.3) show a realistic minimal value.

The measure $\phi_m(bons)$ as defined here can assess the degree of gestural overlap only if we have access to a *dynamically* fixed landmark of the first gesture in question. In a heavily truncated gesture when an interference comes before the dynamically determined peak velocity is reached, the $\phi_m(bons)$ measure cannot be used for the intended purpose as it does not measure the occurrence of the onset of the closing movement with respect to the *independently* acting lip-opening gesture. In such case it simply describes the ratio between time to the occurrence of the disturbance (the onset of the closing gesture in this case) and time to another event (observed but not dynamically determined peak velocity) elicited by the same disturbance. It is possible that the dependence of $\phi_m(bons)$ on *mb-int-norm* reflects precisely this situation and that the gestural overlap in fact continues to increase under increasing time pressure (beyond the fixed phase of peak velocity attainment), a fact that our measure cannot capture.

In the following section, we report on modeling the patterns observed in Sections 3.4 and 3.5. Moreover, in Section 5.2 we explore the validity of $\phi_m(bons)$ measure for the entire interval of overlap degrees as discussed above.

5. Modeling

We present here two models differing in the complexity of their architecture and scope of inter-gestural phasing relations they are designed to investigate. The models are built on the same principles: they search for articulatory patterns that are optimal with respect to competing demands of articulatory efficiency, perceptual clarity and temporal cohesion expressed as a requirement of minimal duration.

These demands are imposed in the form of a composite cost function. The component *E* captures articulatory effort associated with movement of articulators. Parsing cost component *P* reflects the perceptual clarity of the given utterance: the more difficult is the utterance to parse by the listener, the higher the cost. The component *D* assigns cost to duration of the utterance and, as we presently describe, can be modified to influence the relative durations of various parts of the sequence. The overall cost *C* is a linear combination of the three components:

$$C = \alpha_E E + \alpha_P P + \alpha_D D, \tag{5}$$

the less effort, higher clarity, and shorter duration, the lower the cost. These parallel demands usually act in a complementary fashion. Higher clarity as a rule requires greater effort or more time. This complementarity guarantees that there exist optimal solutions: the gestural patterns balancing out the competing influences. The properties of the optimal utterance can be adjusted using the weights α_E, α_P and α_D. Increasing the weight α_D, for example, places higher premium on duration, and pushes the balance towards shorter, faster utterances.

As mentioned earlier, we will explore two models based on the same principles. For modeling the phasing of bilabial gesture /b/ with respect to V_2 in the following section, we deploy the Embodied Task Dynamics (ETD) model designed to search for optimal phasing of gestures acting on a “fully fledged” albeit considerably simplified embodied vocal tract. This model offers a glimpse on the dependence of bV_2 coordination on changing local durational requirements. Due to its design limitations, however, it cannot sufficiently account for inter-gestural phasing relations between subsequent consonantal gestures sharing the same tract variable, as is the case with phasing the lip closing gesture with respect to the preceding lip opening. To investigate this relation, Section 5.2 presents a model based on the same principles that is capable of zooming in on the relationship of this nature.

5.1. Modeling /b/–V2 coordination: Embodied Task Dynamics

The ETD model has been described in detail by Šimko and Cummins (2010, 2011). The inter-gestural timing relations investigated by the model are expressed in the form of a gestural score describing onsets and offsets of gestures participating on a given utterance. The gestural score is realized on a highly simplified model of the vocal tract containing five model articulators: the upper and lower lip, the jaw, the tongue body and the tongue tip. The articulators are inter-connected in a fashion reflecting the anatomical connection in their human vocal tract counterpart: the jaw and the upper lip are linked to an immovable frame of reference (“a skull”), the lower lip to the jaw, and the tongue body to both the jaw and the tongue tip. To simplify computations required by the optimization approach, only vertical movement of the articulators is captured by the model, although the strength of the links between the articulators reflects the horizontal arrangement. Nasalization and voicing are not accounted for by the model.

Only a handful of gestures can be meaningfully simulated by such a stylized model of vocal tract. The model's gestural repertoire consists of a bilabial closure (driving the lip articulators together), an alveolar closure (movement of the tongue tip towards a fixed target) and two vocalic gestures driving the tongue body towards a high (/i/) and low (/a/) position. As no nasalization is implemented we shall use the bilabial gesture to model both /m/ and /b/ in our data; the different labels used in figures below are merely for convenience of the reader.

In a fashion introduced in task dynamical implementation of Articulatory Phonology (Browman & Goldstein, 1992, 1995), each gesture is associated with a second-order critically damped dynamics acting on an appropriate *tract variable*. The bilabial gesture, for example, triggers a dynamics driving a lip aperture variable (the distance between the lips) to its dynamical target set to 0 (collision of the lips). The gestural score prescribes activation intervals of gestures. The ETD model uses a modified pseudo-inversion of the mapping from the model articulator space to that of tract variables to recast the behavior of the latter (defined by gestural score) to articulatory movement (see Saltzman & Munhall, 1989; Šimko & Cummins, 2010).

The ETD model differs from the traditional task dynamical implementation of Articulatory Phonology in three important ways. First, alongside the task-oriented influence of the gestural score, the movement of model articulators is co-determined by their physiological and physical properties – the model vocal tract is embodied. Collisions with oral cavity boundaries and between articulators are accounted for. Second, the articulators are all the time acted upon by a speech-ready dynamics (corresponding to the neutral attractor in AP), linked to a language specific state of the vocal tract suitable for efficient achievement of the goals meaningful for the given language. This dynamics is solely responsible for returning model articulators to their medial positions, including the lips after a successful achievement of a bilabial closure. Consequently, active lip opening gesture is neither necessary nor implemented (purely for reasons of parsimony, its implementation by no means violates the principles of the dynamical modeling considered here). We shall return to this point later.

Finally, in ETD the gestural scores do not serve as a mere input to determine the vocal tract action. They are not derived outside of the embodied system and imposed upon it from ‘above’. Rather, the temporal details of inter-gestural timing are *optimal* solutions of the task to produce a sequence of movements realizing the required utterance. In this sense, the optimal gestural scores are outputs of the optimization process encapsulating the embodied, dynamically controlled model of articulation.

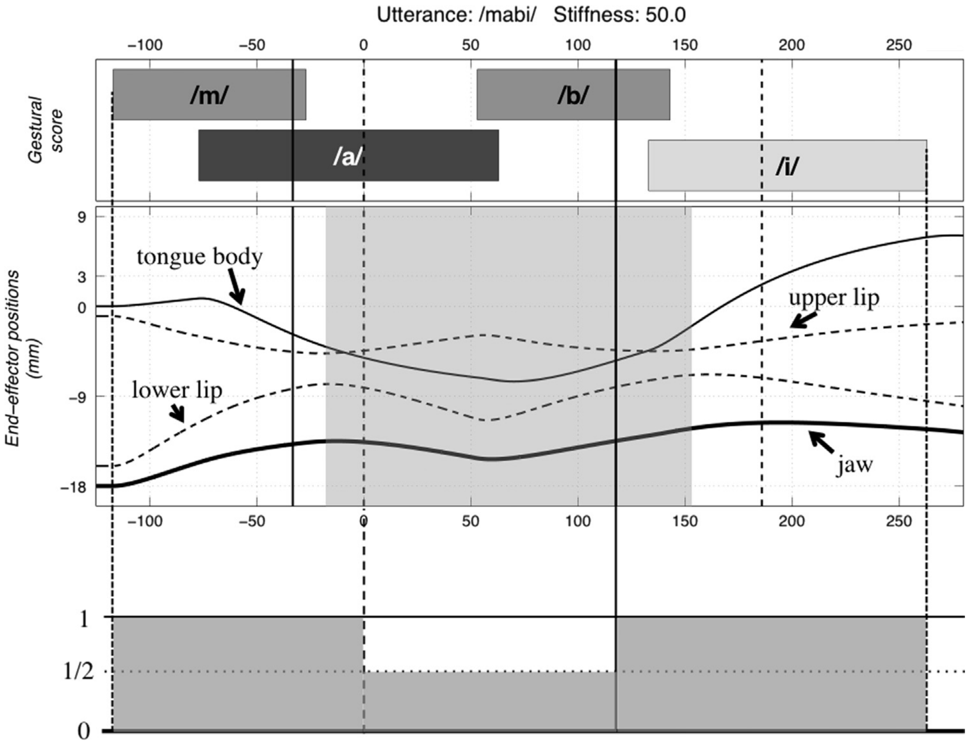

Fig. 8. An example of a (non-optimal) gestural score and articulator trajectories generated by the embodied task dynamics. The temporal extent of *m(♯)b-interval* is shown in the middle panel as the shaded area. The bottom panel illustrates our method of eliciting local relaxation of temporal constraints.

To find an optimal gestural score, three cost functions are defined, as related above. The production cost E associated with articulatory effort is evaluated as overall force expenditure during the realization of the gestural score. The parsing cost P is derived from model's articulation and reflects durations of individual segments (the shorter, the higher cost) and articulatory precision (the less precise articulation the higher cost), see, e.g., O'Dell, Šimko, Nieminen, Vainio, and Lehtinen (2011); Šimko et al. (this issue).[8] The definition of duration cost D, central for the present modeling approach, is described in greater detail below. An optimization procedure is then used to find a gestural score that minimizes the overall cost (5) combining these cost components.

Fig. 8 shows an example of a (non-optimal) gestural score and the corresponding movement of model articulators generated by the ETD model. The activation intervals from the onset of dynamical influence on the appropriate tract variable to its offset are depicted in the top panel. The trajectories in the middle panel are used to evaluate the cost components E and P. The vertical lines show the onsets (solid) and offsets (dashed) of "acoustic closure" for the bilabial gestures /m/ and /b/, respectively, extracted from the articulatory movement (the distance between the model lips reaching a threshold). The gray area shows the extent of m(♯)b-interval defined in the same way as in the data analysis section, i.e., as the interval spanning the maximal constrictions for /m/ and /b/. Note that due to inertia of articulators the beginning and end of this interval do not correspond precisely to gestural offsets of the bilabial gestures.

The duration cost D serves a dual role in the model. First, along with the associated weight α_D (see (5)) it can be used to elicit overall temporal variations of the entire sequence. Second, it acts as a "gestural glue" (cf. Saltzman, Löfqvist, & Mitra, 2000), keeping the gestures together, even overlapping each other, by punishing "idle" intervals of no perceptually relevant action. We use this second property to elicit local temporal variations of interest in this work.

The bottom panel in Fig. 8 illustrates our approach. For "no-break" simulations the cost D is simply computed as the duration of the entire interval under consideration (highlighted on x-axis), i.e., as the area of the rectangle with the height 1. To elicit local temporal relaxing of durational constraints for sequences with a "break", we create a local down-step *temporarily* lowering the height of the rectangular figure. As illustrated by the dotted line in the bottom panel, the figure is lowered to 1/2 during the interval starting at the offset of the closure for /m/ and ending at the onset of /b/-closure. This interval corresponds to the locus of influence of the structural affordance for a break in our recorded material. The duration cost D for this "break" simulation is evaluated as the area of the step-shaped shaded figure. We refer to the height of the step (1/2 in Fig. 8) as a *degree of local binding*, or *DoB*.

As a result, the duration of the interval between "acoustical" closures for /m/ and /b/ counts less than the duration of the rest of the sequence. The gestures participating in determining the interval's duration are under less temporal pressure than the surrounding ones; or, rather, all gestures in the sequence are under less pressure to keep this particular interval short. The lower the degree of binding, the less time pressure on the duration of this interval.

Fig. 9 shows the optimal gestural scores and trajectories for model sequences /m(#abi)/(left) and /m(#iba)/(right) without a "break", i.e. $DoB=1$, (top) and sequences with a "break", $DoB=1/15$ (bottom). As we can see, the interval exposed to the relaxation of time pressure, and consequently also the m(♯)b-interval (shaded in gray in the figure) expanded considerably. More interestingly, this local temporal expansion is accompanied by a change of relative timing between the onsets of /b/ and V_2-gestures in both vocalic contexts: as seen in the optimal gestural scores, in both cases, in "break" simulation the /b/-gesture starts considerably earlier relative the onset of V_2-gesture (/i/ in /m(#abi)/ and /a/ in /m(#iba)/) than in "no-break" case.

[8] Adjustment of the influence of parsing cost conceptually lead to similar hypo- and hyper-variation as manifested in our data. However, in the present paper we do not explore this similarity and focus on localized temporal variations. These are captured via a duration cost component D, central for the present modeling approach, and described in greater detail below.

Please cite this article as: Beňuš, Š., & Šimko, J. Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* (2014), http://dx.doi.org/10.1016/j.wocn.2013.12.005

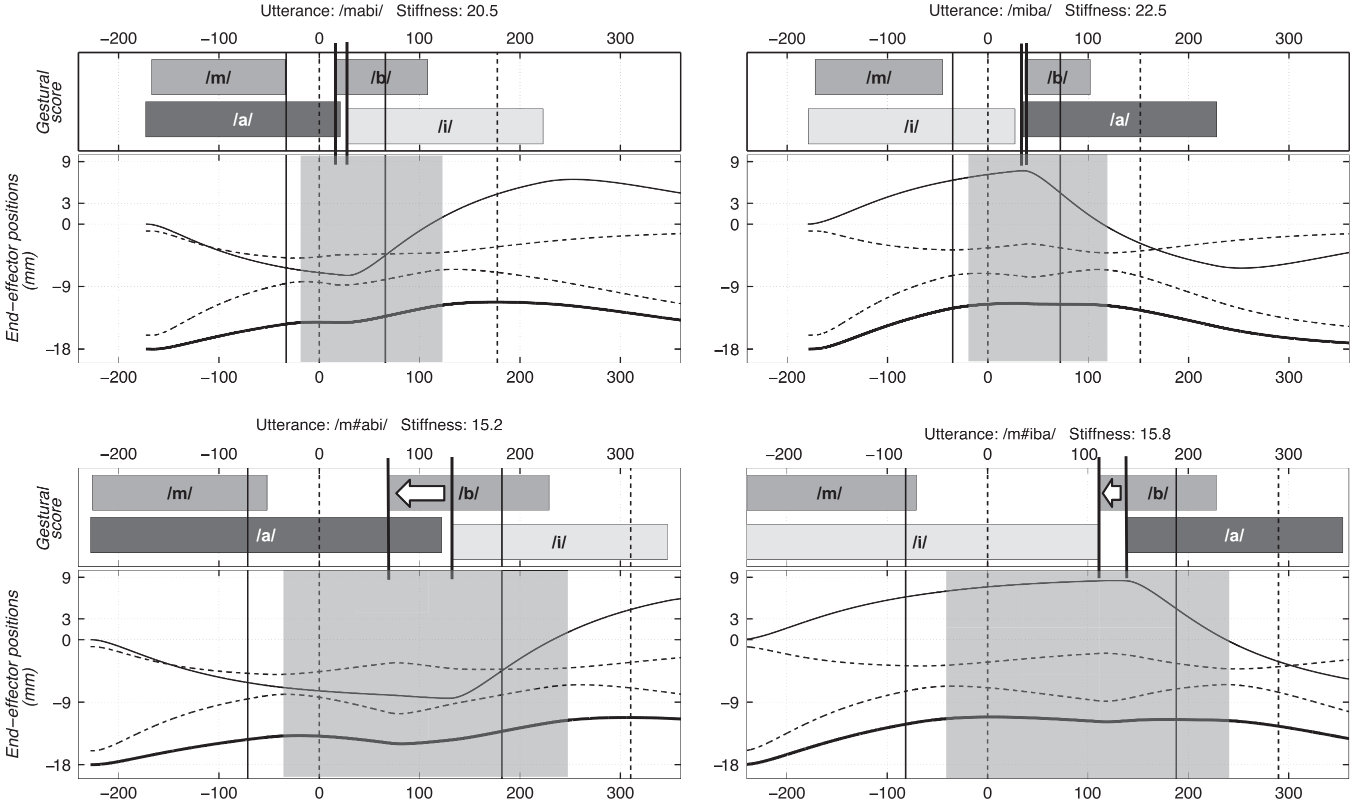

Fig. 9. Optimal gestural scores and articulator trajectories for sequences without (top) and with (bottom) a "break". See text for details.

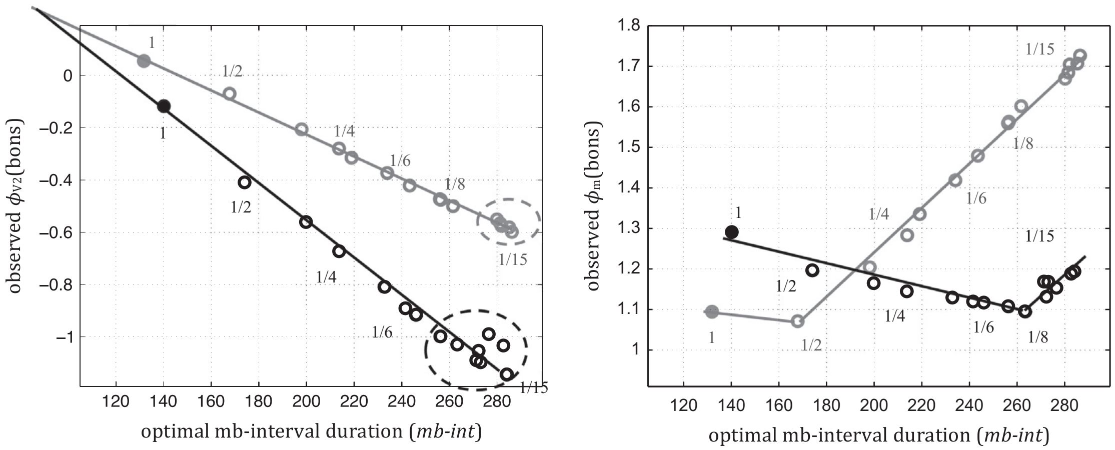

Fig. 10. Relationship between the duration of m(♯)b-interval and phase of the V$_2$-gesture (left) and m-opening gesture (right) at the instant of the onset of lip closing, as predicted by the model. Simulated *abi* sequences are shown in black, *iba* in gray.

To investigate this shift in inter-gestural phasing in more detail, we computed the optimal gestural scores for gradually decreasing values of the degree of binding (for the degrees 1, 1/2, 1/3, …, 1/15) and evaluated the phase of /b/-onset within V$_2$-gesture, ϕ_{V2}(*bons*), and the duration of m(♯)b-interval, *mb-int*, in the same way as for our articulatory recordings.

The left panel in Fig. 10 shows the relation between the two measures as accounted for by our model, the black circles represent /m(#abi)/ and gray ones /m(#iba)/ sequences. The sequences without prosodic break are plotted with full circles. The *DoB*s are also indicated.

As we can see, with a decreasing degree of binding (less local time pressure), the duration of m(♯)b-interval increases and the phase value decreases. That means that with less binding the bilabial closure gesture starts relatively earlier with respect to V$_2$ in sequences, i.e., 'moves to the left'. This behavior of the model reflects our finding reported in Section 3.4 (Fig. 6). As in our data, the slope of the relationship between the m(♯)b-interval duration and phasing is steeper in /m(#abi)/ than in /m(#iba)/ sequences. As the model only includes local but not global temporal variation, there is no need to normalize m(♯)b-interval with respect to speaking rate as we do for our experimental data. Conceptually, the optimal duration of m(♯)b-interval (*mb-int*) as captured by the model corresponds to time-normalized duration (*mb-int-norm*) from the data.

Note also, that for very low binding degree values the lengthening of the m(♯)b-interval and phase shifting stops: the trend does not extend interminably. The data points saturate in a relatively stable area (encircled in the figure). This phenomenon suggests that there is a suitable, sufficient and stable gap (V$_1$-duration) and phasing relation underlying a fully realized prosodic break. From the articulatory point of view, a break is thus not characterized merely by a particular type and value of inter-gestural phasing relation, but by an area of stability thereof.

Please cite this article as: Beňuš, Š., & Šimko, J. Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* (2014), http://dx.doi.org/10.1016/j.wocn.2013.12.005

There are, however, some qualitative differences between the modeling and data analysis results shown in Fig. 6. The (hand drawn) regression lines[9] do not cross: for sequences with no break, the phase is not greater for /m(#abi)/ than /m(#iba)/ as in the data. Even in the most crowded cases, the lip closing does not start relatively earlier in the latter than in the former. While the trends of the lines point in the direction indicating that they would eventually intercept if more crowding was enforced, the model does not account for this. In short, the design of the model, as it stands, does not allow for eliciting crowding influence strong enough to expand the suggested trends to those shown by our data at extremely high speaking rate.

The right hand panel in Fig. 10 shows the dependence between the optimal *mb-int* and the phase $\phi_m(bons)$ of the lip opening movement after /m/ at the onset of /b/ gestures for the same simulation data. The values of these variables were evaluated in the same way as for our recording data and the figure is thus an equivalent of Fig. 7 in Section 3.5. Again, we can observe strong qualitative similarities between the modeling and the data analysis. With relaxing the time pressure, the phase increases; the closing movement shifts further away from the opening movement. For the reasons outlined above (Section 4), the value of opening phase must be greater than 1. In very crowded cases the direction of slope changes suddenly and the (observed) phase starts increasing with increasing binding degree.

To some extent, the model prediction shown in the right panel of Fig. 10 suffers from a similar limitation as the one observed above. In particular for /m(#iba)/ sequences, although the reversal of the trend on the "crowded" side is hinted, it is not very prominent. We hypothesize that the reason for this shortcoming is that the gestural repertoire of the model does not contain an active lip opening, and the lip aperture is not reflected in the evaluation of parsing cost associated with the vowel during which the lip opening occurs. Consequently, the crowding is not associated to the expected degree by increasing the overlap between the opening and closing gestures. Moreover, the ETD modeling as presented here does not give us sufficient insight to the phenomenon of a potential "bifurcation" discussed in Section 3.5.

5.2. Modeling /m/-/b/ coordination

We therefore present a relatively simpler model focusing purely on the lip opening and closing gestures. We ignore the possible interference of lingual movement on the bilabial action, therefore we do not need to include lingual articulators and the jaw in the model. We do not, in fact, include the lips either; the focus of dynamical modeling is purely on the lip aperture (LA) tract variable. The model simulates a succession of lip closing–opening–closing–opening movement representing closing and opening movement for /m/ in our data and subsequent closing and opening for /b/. The aim is to identify the optimal phasing relation between the /m/-opening and following /b/-closing gestures and gestural dynamical parameters that minimize effort and concurrently maximize the extent of lip opening, i.e., increasing the sonority of an intermediate vowel (and decrease the associated parsing cost for the listener). Moreover, we can impose changing temporal constraints, simulating the varying duration of m(#)b-interval.

Behavior of the LA tract variable is again approximated by a critically damped mass spring dynamics. If x represents the value of LA in time t, and x_0 the value of dynamical target for LA, the movement towards this target is described by differential equation in (6) where x', x'' are a velocity and acceleration of x, and k and b are stiffness and damping parameters determining dynamical properties of the movement. As the intended dynamics is critically damped, the correct critical damping value b is analytically linked to stiffness k (b=2 sqrt(k)).

$$x'' = -k(x-x_0)-bx' \tag{6}$$

In the model presented here, the mass is acted upon by four dynamical influences in series. For simplicity, the mass starts in position $x=1$ with velocity $x'=0$. First, a critically damped dynamics with target $x_0=0$ drives the lip aperture downwards, representing lip closing action for /m/ (the dashed gray curve in Fig. 11). When the mass reaches an articulatory target (a "lip closure", this is by necessity greater than the dynamical target and is arbitrarily set to $x_c=0.3$), the first dynamical influence is switched off, and a "lip opening" dynamics takes over. Its dynamical target parameter is $x_0=1$ and stiffness is k_{op}. Black solid curve in Fig. 11 shows the resulting movement of the mass. The opening "gesture" is switched off at a pre-defined phase ϕ_{switch} of its own dynamics. At that moment, a lip closing dynamics kicks in (dashed black curve in Fig. 11), with stiffness k_{cl} and target $x_0=0$. Once again, this influence switches off when the target $x_c=0.3$ is reached. Finally, the "lips" are driven apart again by dynamics with the target value $x_0=1$ (solid gray curve in Fig. 11). This dynamics is switched off again at an arbitrary point when the minimum of LA during "/b/-closure" can be identified. Parsimoniously, the stiffness parameters of the initial closing and final opening gestures are set to the same values as their counterparts in the middle of sequence, i.e., k_{cl} and k_{op}.

The movement pattern of the mass is determined by three parameters only: the stiffness values k_{cl} and k_{op} and the phase of the opening gesture when the closing movement starts ϕ_{switch}. To find "appropriate" values of these parameters is the task for an optimization process.

Each movement sequence is assigned a cost. The cost function, mapping the parameters k_{cl}, k_{op} and ϕ_{switch} to a cost value combines three components. The effort E is approximated through the stiffness parameters: the stiffer the gestures the more force driving the lips together and subsequently the more effort required. It is defined as $E=(2k_{cl}+k_{op})^2$ eliciting slower opening gestures compared to their closing counterparts (this choice does not influence the qualitative predictions below and was made to reflect general tendency of closing gestures to be faster than the opening ones).

The parsing cost P is related to the maximal opening (maximal value of x) reached between the two closures: the smaller the lip aperture the less sonorant the medial vowel and therefore the greater demands on the listener. It is evaluated as the reciprocal of the difference between the LA value at

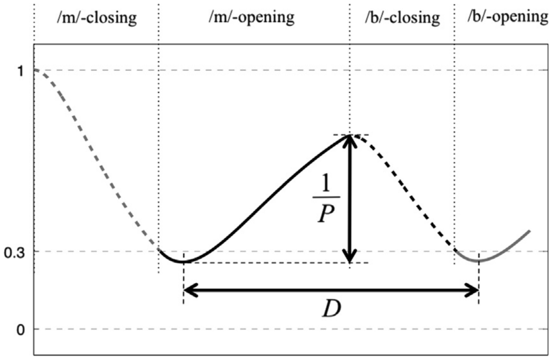

Fig. 11. An illustration of the effects of successively triggered dynamical influences representing lip closing and opening for /m/ and lip closing and opening for /b/.

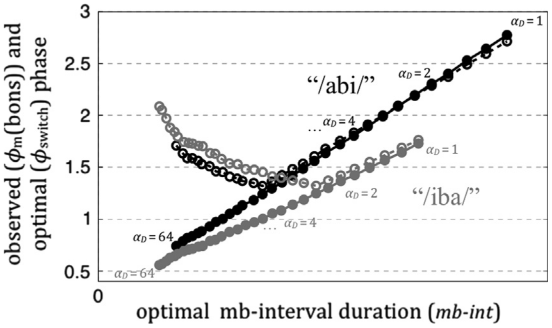

Fig. 12. Relationship between the duration of m(♯)b-interval and the phase (observed and optimal) of m-opening gesture at the onset of lip closing in /a-i/ (black) and /i-a/ context (gray), as predicted by the model.

the time of maximal lip closing for /m/ and its value at the point of maximal opening (see Fig. 11). Finally, as shown in Fig. 11, the duration cost *D* is simply the duration of the interval between maximal lip constrictions for /m/ and /b/.

The relative influence of these cost components on the overall cost *C* defined in the same way as before (5) is controlled by the weights α_E, α_P and α_D. Greater α_P places greater premium on the lip opening during the interval between the constrictions. We use this for distinguishing between medial /a/ and /i/, assuming that the lip opening matters more for the vowel /a/ than for /i/. Consequently, we set $\alpha_P = 1$ for /i/ context (our sequences /am(#)iba/) and $\alpha_P = 3$ for /a/ (sequences /am(#)abi/). Everything else being equal, the optimal sequences with the /a/ settings can be expected to reach greater maximal lip opening than /i/-sequences.

Temporal pressure can be elicited through the duration weight α_D. conceptually equivalent to *DoB*. The greater its value, the more costly the longer sequences become, resulting in shorter (and faster with less opening) optimal movement patterns. Note that the durational cost component is conceptualized in a slightly simplified way compared to the “fully-fledged” ETD model used in the previous section: the influence of the *DoB* parameter spans the articulatory m(♯)b-interval (as defined in Section 2.5) rather than the interval between the “acoustical” closures of /m/ and /b/ as before. The reasons for this discrepancy are technical and do not impact qualitative nature of the results.

The third weight is kept constant in the simulations ($\alpha_E = 1$) as it does not influence the *relative* contributions of precision and duration demands to the optimal patterns.

To find optimal phasing relations between the opening and closing bilabial gestures, we searched for the parameters minimizing the overall cost. We modeled the decreasing time pressure (providing an opportunity for a break) by gradually lowering the weight α_D. For each optimal sequence, we computed the phase of the m-opening gesture ϕ_m(*bons*) at the point when the b-onset occurred in the same way as for our data, that is from the *observed* movement traces as shown in Fig. 11 (rather than directly from the dynamical description at our disposal, namely ϕ_{switch}). Similarly, we evaluated the observed m(♯)b-interval duration, *mb-int.*

Fig. 12 summarizes the simulations results. The optimal durations of m(♯)b-interval and corresponding values of the optimal phase ϕ_{switch} (full circles) and observed phase ϕ_m(*bons*) (empty circles) are plotted in black for “/a/” setting ($\alpha_P = 3$) and in gray for “/i/” setting ($\alpha_P = 1$). The time pressure values representing by α_D range from 1 to 64 (with a step increasing exponentially) and are shown next to the representations of the optimal constellations.

The figure offers several observations. As expected, with the growing time pressure the optimal duration of the m(♯)b-interval decreases. At the same time, the optimal phase ϕ_{switch} of the lip opening gesture at which the closing movement starts decreases (full circles). It is important to note, that ϕ_{switch} is the *real* phase of the opening gesture at which the closing movement starts prescribed by the optimal coordination patterns, and as such it is not subject to computational limitations outlined in Section 4.

The empty circles in Fig. 12, on the other hand, show the *observed* value of phase ϕ_m(*bons*) computed in the same way as in our data. For the reasons highlighted earlier, its value must be greater than 1. Fig. 12 corroborates this fact. As the real phase decreases, the estimate follows its values closely up to a point when the onset of the closing movement gets close to the instant when the lip opening movement would naturally reach its peak velocity determined by its dynamics. From this point onwards (to the left in our figure), the observed estimate abruptly diverges from the real setting. For greatly overlapped gestures, the observed phase does not capture the dynamical phasing relationship between them.

Comparing Fig. 12 with Fig. 7 in Section 3.5 reveals deep similarities. The apparent abrupt change of behavior shown by modal voice compared to utterances with a prosodic break in Fig. 7 could indeed be likely ascribed to the reasons described above: rather than signaling a different phasing relationship for crowded gestures it reflects limitations of phase estimation.

Of course, a mere similarity of figures is not a proof of this conjecture but there are two facts giving this explanation a strong standing. First is a methodological one known as Occam's razor: the hypothesis of continuation of the trend (increasing overlap under time pressure) is a simpler one explaining the data as it does not require an additional assumption of an abrupt change. The second reason is that this behavior is, as our modeling suggests, an optimal one under competing requirements of production, perception and temporal efficiency.

Our simulations match the data in another important aspect. Note that the mean of sampled durations of m(♯)b-interval is smaller for /iba/ than for /abi/: the smaller demands on the lip opening for /i/ result in shorter duration interval for this vowel under similar time pressure. This is in line with traditional explanation for generally shorter acoustic duration of the vowel /i/ compared to /a/ (e.g. Catford, 1977). Consequently, both in the optimal sequences and in data analysis results, the slope of the relationship between the m(♯)b-interval and phase is steeper for /abi/ than for /iba/. This suggests greater sensitivity of phasing relation between the opening and closing movement when the underlying vowel is /a/ for which the lip opening plays more important role than for /i/.

[9] We didn't attempt statistical regression for two reasons: (1) the data points are generated by quasi-deterministic process and are not subject to the same stochastic influences as our recordings and (2) the slopes are influenced by the data points in the stable area and by our arbitrary decision how far to relax the *DoB* value.

6. General discussion

We employed a bottom-up strategy for eliciting prosodic boundaries in which the linguistic structure provided only an affordance for the break to optionally, and unintentionally, emerge as a potential strategy for resolving low-level continuous prosody demands of rate and precision. Our data thus complement more traditional top-down approaches to investigating prosodic variation involving prosodic contrasts, often induced by discrete syntactic/pragmatic contrasts.

After establishing a close positive relationship between the local slow down (*mb-int-norm*) and increased salience of prosodic breaks signaled with deviations from modal voice setting, we investigated the phasing of the onset of the lip closing gesture towards a constriction and the following post-boundary vocalic transition. Data analysis and modeling suggest that phasing relations among gestures in the vicinity of structural affordance for a phrasal boundary undergo quantitative changes. As the localized temporal pressure decreases, the temporal coordinations of the surrounding gestures get rearranged. The gestures become less “crowded.”

The metaphor of “crowding” describes the patterns revealed by the phase analysis quite well. In the tokens with no discernible prosodic boundary, the onset of b-gesture was “squeezed” between the canonically surrounding gestures: it was triggered at the early phase of the lip opening movement of the preceding /m/ and occurred at a similarly early phase of the vocalic transition towards the following vowel. As the pressure gradually released, the b-onset moved further away from both surrounding gestures.[10]

These consequences of changes in boundary strength found in the data correspond well with the results of modeling the emergence of phrasal boundary using an embodied optimization approach. The Embodied Task Dynamics model (ETD), simulating the local temporal pressure by gradually varying demands of temporal cohesion, predicts qualitatively equivalent patterns, in particular for the b-onset–V_2 relationship. Predictions of similarly motivated model of coordination of two adjacent homorganic consonantal gestures elucidated the possible emergence of gradual relaxation of phasing between the opening and closing bilabial gestures with decreasing temporal pressure, which was only suggested by the “fully-fledged” ETD.

The modeling results are largely consistent with the prosody-driven bV_2-phasing pattern found in our data. Although the optimal phases for /m(#)abi/ and /m(#)iba/ sequences did not reach the observed reversal in relation to vocalic context, the trends towards it were present. Moreover, when the boundary strength, modeled as a degree of binding, weakened beyond a certain threshold, the bV_2 phasing relation stabilized. The agreement between the data analysis and the model suggests that alleviating the pressure by prosodic boundary (weakening gestural crowding) corresponds asymptotically to the optimal phasing relations in sequences without an interference of homorganicity. As the boundary emerges, the crowds disperse.

Consequently, an important finding of our paper is a methodological one. Most studies investigating the articulatory realization of prosodic structure, ours included, use a succession of bilabial consonants interspersed with vowels for the target sequences. The rationale is that the effects of consonants on vowels (and vice versa) should be minimal to allow for reliable identifications of movements pertaining to individual segments and thus clearer effects of prosody on their kinematics. We observe, however, that gestural crowding resulting from a succession of homorganic consonants interacts with prosodic demands in non-trivial ways. Our data suggest that only for very strong breaks, the factor of homo/hetero-organic nature of the consonant preceding the target CV-phasing might play an insignificant role but with decreasing boundary strength, this factor enters into the demands for efficient productions. Since the realization is assumed to arise as a successful resolution of these demands, this factor should be taken into consideration when prosody and phasing are investigated. We do not mean to suggest that homorganic consonants should be avoided in studies investigating temporal sequencing of gestures and prosodic structure. After all, the upshot of employing successive bilabials and limiting CV coarticulation in target CVCV sequences is well motivated. Rather, we suggest that: (1) generalizations of the already reported data on phasing relations involving homorganic consonants, especially for weaker boundary strengths, might not be straightforwardly extendable to other phonetic contexts, and (2) exploring differences in observed phasing in data that systematically vary homo-/hetero-organic nature of adjacent consonants is needed.

Both versions of optimization modeling used in this work suggested a continuous nature of the effects discussed above. Correspondingly, the data analysis did not reveal any obvious discontinuities in the patterns. The occurrence and the strength of prosodic breaks coincided with the release of local temporal pressure represented by the lengthening of cross-boundary m(♯)b-interval. The tokens were continuously more ‘boundary-like’ in relation to the temporal overlap of the pre- and post-boundary lip closures. This is supported by the trends in our data relating the local temporal pressure to both vocal folds activity and phasing between articulatory movements in the vicinity of the boundary. Moreover, given that our data probe boundary strengths biased toward weak and even non-existing breaks (re-syllabification was common), we explored how trends in our data correspond to results obtained using discrete elicitation paradigms. An extrapolation of trends from our data shows a considerable qualitative agreement with the findings based on stronger discretely elicited breaks discussed in the introduction.

Continuity is the prevailing feature of our supraglottal articulatory data despite the emergence of discrete-like variability of boundary strengths associated with the glottal activity. Interestingly, even the seemingly discrete behavior of vocal folds setting – provisionally discretized in our approach to modal voice, glottalization and silence – shows continuous features in smooth trends of temporal patterning (e.g. Figs. 6 and 7) both within and across these three categories. This continuous correlation between local time pressure and the occurrence and extent of the vocal folds closing gesture suggests a lawful relationship between the two. This is supported by the saturation of the optimal solutions at the end of the continuum showed in the left panel of Fig. 10. This suggests that an optimal realization of a prosodic boundary beyond certain critical value of a local slow-down, requires a different articulatory-perceptual mechanism for increasing the boundary strength, which might be provided by the interplay of glottal and supraglottal gestures. Additionally, glottal closure, and acoustic silence resulting from it, increases the perceptual saliency of a prosodic boundary.

Of course, the fact that we found no discontinuities in the articulatory marking of prosodic chunking in our data should not stop us from searching for the dynamic model representing the cognitive system relating continuous and discontinuous aspects of prosody. In fact, the observation that discrete behavior (presence of a prosodic break) spontaneously emerges from the variation in our continuous parameter provides support for this direction of future research. Non-linear dynamics might provide a suitable formal framework for capturing and explaining the studied phenomena (Gafos, 2006; Gafos & Beňuš, 2006, reviewed in Section 1.3). If the vocal cord activity is conceptualized as an order parameter with stable attractors corresponding to modal voice and silence, the normalized m(♯)b-interval can be taken as a control parameter: continuously decreasing premium on local time results in a switch from one stable mode (modal voice) to another (silence) through an unstable region of glottalization. Hence, a parameter

[10] The metaphor of “crowding” might evoke somewhat negative connotations. Therefore, we hasten to say that we do not by any means suggest that homorganic gestures are dispreferred and should be avoided. All our results indicate is that homorganicity of proximal articulatory gestures presents different constraints on inter-gestural phasing to those evoked by heterorganic gestures. Although speakers routinely deal with each type of constraint in an efficient manner, researchers should be aware of this phenomenon when they interpret their results.

expressing the local tightening–relaxing of temporal cohesion at the structural affordance for break might provide a lawful relationship to both supraglottal phasing in the vicinity of the break as well as for more granular glottal setting.

Our modeling approach introduces a local relaxation of inter-gestural temporal cohesion in the vicinity of a prosodic boundary. In this sense it is similar to the influential account of Byrd and Saltzman (2003) postulating prosodic layer of π-gestures that locally slow down activation functions of articulatory gestures. In fact, we fully embrace their idea of "(v)iewing the boundaries as warping the temporal fabric of an utterance" (p. 176). It is actually plausible, that the patterns revealed in our data analysis could be successfully modeled using an ensemble of π-gestures or more general temporal modulation gestures (Saltzman et al., 2008). As stated earlier, the main difference between the two approaches lies in motivation. While in π-gesture theory the changes in inter-gestural coordination come as a result of an additional control layer, in the ETD model they emerge as an optimal resolution of localized communicative intentions. The qualitative agreement between the predictions of the ETD model and the results of data analysis indicates that the speakers realize their intention to signal a boundary in the way that satisfies efficiency requirements imposed by embodied nature of speech production and perception systems (cf. Lindblom's (1999) program of Emergent Phonology).

Considering further the relevance of our approach to the conceptualization (and modeling) of boundary strength in particular and prosodic structure in general, the role of degree of binding (*DoB*) parameter should be discussed. *DoB* represents a local temporal constraint placed on the m(♯)b-interval; the word "binding" refers to a relationship between the two consecutive bilabial gestures for /m/ and /b/. The greater its value, the higher the premium placed on temporal cohesion between these gestures; consequently (and trivially) the shorter the m(♯)b-interval relative to the surrounding speech. The effect of changing the *DoB* corresponds to elicitation of variation in relative duration of the m(♯)b-interval considered in our data analysis. As both modeling and data analysis suggest, the relative phase used as a measure of inter-gestural coordination varies continuously as a function of the relative duration of the m(♯)b-interval.

Regarding the planning/linguistic level, we tentatively suggest that it is the local tightening–relaxing of temporal cohesion constraints (represented by *DoB* parameter) that might be a suitable candidate for a high-level control parameter simultaneously driving the changes in m(♯)b-interval duration and in inter-gestural phasing. The "value" of this parameter arises from communication demands of a type "how strong a break, if any, do I want/need to produce at this point so that it gets appropriately interpreted by my audience." This is the planning/linguistic level that seems to be sufficient to elucidate phenomena observed in our data. We are by no means opposed to a possible discretization of this cohesion parameter leading to several discrete boundary types identified in literature. A future direct comparison of data with breaks emerging through low-level phonetic variations (like ours) and planned breaks expressing linguistic contrasts will test the hypothesis that both species of prosodic breaks (planned and emergent ones) can result from the same mechanism, and be implemented in the same way. Our data provide two partial bits of support for this hypothesis. First, our results with coordination patterns show continuity both within our discrete categories (MV, Glot, Sil), as well as across them, and, extrapolating these patterns to the extremities quite closely correspond to reported differences between non-break and break in data elicited with more traditional paradigms based on linguistic planning. Second, we reported similarity of patterns for raw m(♯)b-interval and normalized one. We assume that the raw measure primarily reflects general tempo variability and the normalized one taps more directly into the prosodic structure. The observed similarity thus suggests that local (linguistic) and global (phonetic) temporal variations exert qualitatively similar influence on inter-gestural coordination.

Of course, the hypothesis above does not exclude the option, which we consider highly plausible, that there exist discrete-like "clusters" of favored production patterns in various languages (including Slovak), and that these clusters, modeled with, for example, values of locally induced temporal variation formalized as Degree of Binding in our model, get automated in shape of discrete linguistic motor plans during speech acquisition process. This may have lead to phylogenetic "phonologization" of the boundary strength. Our hypothesis, however, predicts that these clusters – or, more precisely, the articulatory realizations in terms of inter-gestural coordination – lie along a continuum as discussed in this work.

We are aware of a methodological limitation concerning the relationship between the data and predictions of the models. We reported that the patterns observed in the data fit the model's predictions. It, however, does not necessarily imply that the underlying mechanisms driving speech production are identical to those eliciting the similar patterns in the model. This limitation is inherent for a modeling approach in general. The considerable agreement between the empirical findings and modeling results, however, provides a strong support for embodied optimization-based dynamical paradigm used here. It shows that the platform originally developed for exploring fine details of inter-gestural phasing can be extended to account for higher-level prosodic characteristics of speech (partly because, as this work suggests, the emergence of prosodic phenomena may be accompanied by lawful continuous changes in phasing). The fact that the behavior of the speakers to a considerable degree agreed with the predictions – including complex interactions between the strength of the prosodic boundary, induced by low-level precision/tempo demands, and context-dependency of inter-gestural phasing – provides a strong support to the hypothesis that speech production, as many other types of embodied skilled action, is to a large extent shaped by the principles of optimality.

7. Conclusions

This paper examined the relationship between emergent prosodic boundaries in *am*(#)*iba* and *im*(#)*abi* sequences and gestural phasing of cross-boundary /m-b/ and post-boundary bV sequences. The novel aspects of the approach include: (1) employing continuous low-level variation in tempo and articulatory precision for inducing high-level prosodic break emergence, (2) combining this type of prosodic variation with asymmetric vocalic environment, (3) relating supraglottal gestural phasing with the vocal cord activity at the structural affordance for a prosodic boundary, and (4) extension of the embodied task-dynamics approach that models boundary strength as gradual releasing of demands for temporal cohesion among gestures localized to this structural affordance. The patterns observed empirically in our data, in which prosodic breaks emerge from low-level tempo/precision variation, when extrapolated to high boundary strengths, are similar to the patterns reported in the literature and elicited through discrete high-level prosodic variation. The results of data analysis show qualitative agreement with the predictions of the extended ETD model. Based on this agreement we argued that the prosodically-driven re-arrangements of inter-gestural phasing might arise as a result of efficient resolution of tradeoffs among articulatory effort, perceptual clarity, and localized adjustments of temporal cohesion.

Acknowledgments

This research was supported by Alexander von Humboldt Grants to both authors, and the preparation of the manuscript also by the Slovak Ministry of Education under VEGA no. 2/0202/11 Grant and project RPKOM, ITMS 26240220064 funded by the Research & Development Operational

Please cite this article as: Beňuš, Š., & Šimko, J. Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* (2014), http://dx.doi.org/10.1016/j.wocn.2013.12.005

Programme funded by the ERDF. The authors wish to thank two anonymous reviewers and the associate editor for valuable comments to the earlier drafts of this paper and Fred Cummins for inspiration and assistance with the data collection. All mistakes are the authors' responsibility.

References

Bachenko, J., & Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, *16*(3), 155–170.

Beckman, M.E., & Edwards, J. (1992). Intonational categories and the articulatory control of duration. In: Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 359–375). Tokyo: Ohmsha.

Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In: G. Docherty, & D. R. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 68–86). Cambridge: Cambridge University Press.

Benoit, C. (1986). A note on the use of correlation in speech timing. *Journal of the Acoustical Society of America*, *80*, 1846–1849.

Beňuš, Š. (2012). Phonetic variation in Slovak yer and non-yer vowels. *Journal of Phonetics*, *40*(3), 535–549.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155–180.

Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In: R. F. Port, & T. van Gelder (Eds.), *Mind as motion* (pp. 175–194). Cambridge, MA: MIT Press.

Brunner, J., & Zygis, M. (2011). Why do glottal stops and low vowels like each other? In *Proceedings of ICPhS XVII*. Hongkong.

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, *57*, 3–16.

Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000). Phrasal signatures in articulation. In: M. Broe, & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 70–87). Cambridge: Cambridge University Press.

Byrd, D., Krivokapić, J., & Lee, S. (2006). How far, how long: On the temporal scope of prosodic boundary effects. *Journal of the Acoustical Society of America*, *120*, 1589–1599.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*, 173–200.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *3*, 149–180.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press.

Cho, T. (2002). *The effects of prosody on articulation in English*. New York and London: Routledge.

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *32*, 141–176.

Cho, T. (2006). Manifestation of prosodic structure in articulation: Evidence from lip movement kinematics in English. In: L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory phonology 8: Varieties of phonological competence* (pp. 519–548). New York: Walter De Gruyter.

Cho, T., Lee, Y., & Kim, S. (2011). Communicatively driven versus prosodically driven hyper-articulation in Korean. *Journal of Phonetics*, *39*(3), 344–361.

Cummins, F. (1999). Some lengthening factors in English speech combine additively at most rates. *Journal of the Acoustical Society of America*, *105*(1), 476–480.

De Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, *96*(4), 2037–2048.

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.

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3740.

Gafos, A. (2006). Dynamics in grammar: Comment on Ladd and Ernestus & Baayen. In: L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory phonology 8: Varieties of phonological competence* (pp. 51–79). New York: Mouton de Gruyter.

Gafos, A., & Beňuš, S. (2006). Dynamics of phonological cognition. *Cognitive Science*, *30*, 905–943.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Hoole, P., & Zierdt, A. (2010). Five-dimensional articulography. In: B. Maassen, & P. H. H. M. van Liehout (Eds.), *Speech motor control* (pp. 331–349). Oxford: OUP.

Kelso J. A. S., & Tuller, B. (1985). Intrinsic time in speech production: Theory, methodology, and preliminary observations. In *Haskins laboratories status report on speech research* (pp. 23–39), Vol. SR-81. Haskins Laboratories.

Krivokapić, J. (2007). *The planning, production, and perception of prosodic structure [unpublished Ph.D. thesis]*. University of Southern California.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In: W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer Academic Publishers.

Lindblom, B. (1999). Emergent phonology. In *Proceedings of 25th annual meeting of the Berkeley Linguistics Society*. Berkeley: University of California.

Löfqvist, A., & Gracco, V.L (1999). Interarticulator programming in VCV sequences: Lip and tongue movements. *Journal of the Acoustical Society of America*, *105*(3), 1864–1876.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris.

Nittrouer, S. (1991). Phase relations of jaw and tongue tip gestures in the production of VCV utterances. *Journal of the Acoustical Society of America*, *90*, 1806–1815.

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America*, *84*, 1653–1661.

O'Dell, M., Šimko, J., Nieminen, T., Vainio, M., & Lehtinen, M. (2011). Relative timing of bilabial gesture in Finnish. In *Proceedings of ICPhS XVII*. Hongkong.

Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. (2002). Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America*, *112*, 1627–1641.

Pierrehumbert, J., & Frisch, S. (1994). Source allophony and speech synthesis. In *Proceedings of the ESCA/IEEE workshop on speech synthesis* (pp. 1–4).

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In: G. Doherty, & D. R. Ladd (Eds.), *Papers in laboratory phonology II: Gesture segment prosody* (pp. 90–117). Cambridge: Cambridge University Press.

Saltzman, E. L., Löfqvist, A., & Mitra, S. (2000). 'glue' and 'clocks': Intergestural cohesion and global timing. In: M. Broe, & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 88–101). London: Cambridge University Press.

Saltzman, E., & Munhall, 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, and C. Reis (Eds.), *Proceedings of speech prosody* (pp. 175–84).

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371–405.

Šimko, J., & Cummins, F. (2010). Embodied task dynamics. *Psychological Review*, *117*(4), 1229–1246.

Šimko, J., & Cummins, F. (2011). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science*, *35*(3), 527–562.

Šimko, J., Cummins, F., & Beňuš, Š. (2011). An analysis of the relative timing of coarticulated gestures within VCV sequences. In *Proceedings of ICPhS XVII* (pp. 1850–1853). Hong Kong.

Šimko, J., O'Dell, M., & Vainio, M. Emergent consonantal quantity contrast and context-dependence of gestural phasing. Journal of Phonetics, http://dx.doi.org/10.1016/j.wocn.2013.11.006, this issue.

Smiljanic, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Linguistics and Language Compass*, *3*(1), 236–264.

Truckenbrodt, H. (1999). On the relation between syntactic phrases and phonological phrases. *Linguistic Inquiry*, *30*, 219–255.

Please cite this article as: Beňuš, Š., & Šimko, J. Emergence of prosodic boundary: Continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics* (2014), http://dx.doi.org/10.1016/j.wocn.2013.12.005