Emergent consonantal quantity contrast and context-dependence of gestural phasing

Juraj Šimko[a,b]*, Michael O'Dell[c], Martti Vainio[b]

*Corresponding author: juraj.simko@helsinki.fi
[a] Bielefeld University, Univestitätsstraße 25, 33615 Bielefeld, Germany
[b] University of Helsinki, Siltavuorenpenger 3 A, 00014 University of Helsinki, Finland
[c] University of Tampere, Kanslerinrinne 1, 33014 University of Tampere, Finland

Abstract
Embodied task dynamics is a modeling platform combining task dynamical implementation of articulatory phonology with an optimization approach based on adjustable trade-offs between production efficiency and perception efficacy. Within this platform we model a consonantal quantity contrast in bilabial stops as emerging from local adjustment of demands on relative prominence of the consonantal gesture conceptualized in terms of closure duration. The contrast is manifested in the form of two distinct, stable inter-gestural coordination patterns characterized by quantitative differences in relative phasing between the consonant and the coproduced vocalic gesture. Furthermore, the model generates a set of qualitative predictions regarding dependence of kinematic characteristics and inter-gestural coordination on consonant quantity and gestural context. To evaluate these predictions, we collected articulatory data for Finnish speakers uttering singletons and geminates in the same context as explored by the model. Statistical analysis of the data shows strong agreement with model predictions. This result provides support for the hypothesis that speech articulation is guided by efficiency principles that underlie many other types of embodied skilled action.

Keywords
dynamics, optimization, geminates, Finnish, production, perception, context-dependency

Emergent consonantal quantity contrast and context-dependence of gestural phasing

1.0 Introduction

The phonological contrast between singleton and geminate stop consonants is phonetically realized primarily through a difference in the duration of oral closure. In this paper we investigate temporal and kinematic articulatory characteristics of lip movement during the production of short and long voiceless bilabial stops in Finnish. We focus on the coordination patterns of the labial movement with tongue gestures associated with the flanking vowels. In particular, we explore the lawful variation of these temporal and kinematic patterns in relation to articulatory properties of gestures in the vicinity of the bilabial stop. We present the results in the context of predictions of an optimization based dynamical modeling paradigm which interprets articulatory patterns underlying the consonantal quantity contrast as distinct stable solutions of an optimization task, representing distinct modes of coordination. Our modeling and experimental results suggest that a phonological contrast can emerge through discretization of the continuous space of inter-articulator coordination, arising from complementary requirements of efficient and effective communication.

The obvious characteristic distinguishing short and long consonants is the difference in constriction duration. The reported ratio between duration of long and short non-stop consonant is highly language, speaker and consonant dependent. It ranges from about 1:1.1–1:1.4 for Norwegian non-stop consonants (Fintoft, 1961), through 1:1.65–2.35 for Italian (Kingston *et al.*, 2009) and 1:2–1:3 for Finnish (Lehtonen, 1970) to as much as 1:4 for Japanese voiced alveolar stops (Homma, 1981).

Relative robustness of the durational difference may be related to additional cues signaling quantity contrast. In some languages, e.g., Swedish or Italian, where the duration of the consonant itself may not be sufficient for marking the contrast, the constriction lengthening in long consonants is accompanied by shortening of the vowel immediately preceding the consonant (Elert, 1964; Esposito & Di Benedetto, 1999). This shortening phenomenon is, however, not universal (cf. Ridouane, 2010). Finnish and Japanese speakers, for example, actually often *lengthen* the vowels before geminates (Lehtonen, 1970; Port *et al.*, 1987).

A modeling approach aimed at explaining the differences between languages has been implemented by Smith (1995) within the powerful framework of Articulatory Phonology (Browman & Goldstein, 1992). Her model accounted for the observed differences between languages through qualitatively different organization of inter-gestural coupling relations. Another hypothetical explanation of similarities and differences among languages is based on principles of the theory of "adaptive dispersion" proposed by Lindblom (1987), where production is guided by a principle of sufficient contrast presented by the speaker to a listener (Engstrand & Krull, 1994).

Our approach to modeling quantity, although not focused on differences between languages, combines elements of the two approaches outlined in the previous paragraph. As our modeling platform is purely articulatory, we now turn our attention to a body of research investigating articulatory aspects of quantity contrast.

As expected, articulatory measurements mirror acoustic analysis and show significantly longer durations of articulatory closure for geminates compared to singletons (Bouarourou *et al.*, 2011, for Tarifit Berber; Zeroual *et al*., 2008, for Moroccan Arabic; Löfqvist, 2005, 2006, 2007, for Japenese, O'Dell *et al.*, 2011b, for Finnish). Consequently, movements of relevant articulators for geminates have longer durations than for singletons. Zeroual *et al*. (2008), for example, found that for voiceless and voiced alveolar stops in Moroccan Arabic the duration of the entire tongue tip gesture (from the onset of the movement towards articulatory constriction to the end of the constriction release movement) is significantly greater for geminates than for singletons.

Their measurements also show significantly greater durations for geminates of the movement towards the *maximal constriction* and of the release movement from maximal constriction to the end of the movement. Interestingly, the sub-intervals from the movement onset to the target achievement as well as from the release from the constriction target to the end of the movement were not significantly influenced by consonantal quantity (Zeroual *et al*., 2008); the durational differences mentioned above presumably arise during the constriction interval.

There are also some spatial differences between singleton and geminate articulation, in the expected direction. Löfqvist (2005) reports that the lower lip reached a significantly higher position for geminates than for singleton bilabials in Japanese. A similar effect of quantity has been found for Finnish (O'Dell *et al.*, 2011b). Amplitude of the opening movement of the lower lip was found to be greater for geminates than singletons in Italian bilabial consonants (Gili Fivela & Zmarich, 2005).

The difference in velocity of relevant articulators during closing movement towards constriction does not show a universal pattern distinguishing geminates form singletons. Smith (1995) reports lower articulatory velocity in geminate gestures compared to singletons for Japanese as well as for two out for three Italian speakers she investigated. Japanese speakers analyzed by Löfqvist (2005) also tended to realize geminates with lower peak velocity of the upper lip, but a Swedish speakers in the same study used significantly *higher* upper lip peak velocity in geminates. No significant effect of quantity on the peak velocity was found for the tongue tip closing gesture in the Morrocan Arabic study of Zeroual *et al*. (2008). Measurements in Bouarourou *et al.* (2008) for Tarifit Berber suggest a higher articulator velocity for geminates at the instant of oral closure (velocity at closure can be expected to correlate strongly with peak velocity value, cf. Löfqvist & Gracco, 1997), but Gili Fivela *et al.* (2007) found no evidence of a similar quantity-related effect in their Italian data.

Singleton and geminate consonants are, in general, co-produced with an articulatory transition between the flanking vowels. The duration of this transition, measured as tongue body movement between the attainment of the articulatory targets for the vowel preceding and following the consonant, tends to be longer when the consonant is geminate rather than singleton. A reliable difference in this direction has been identified for Japanese (Smith, 1995; Löfqvist, 2006, with several exceptions in words /kema/ and /kemma/) and for vowel related lip rounding transition for Finnish (Lehtonen, 1979). Gili Fivela *et al*. (2007) reported similar lengthening of vowel-to-vowel interval duration for Italian. Smith (1995), however, reported no effect of the consonant quantity on the intervocalic transition for one Italian speaker, and a significant *shortening* of the interval for another two Italian speakers.

Löfqvist (2006, 2007) found no significant influence of consonantal quantity on the spatial extent of the lingual movement between the flanking vowels when the consonant was bilabial, but the movement was generally larger for lingual geminates. Consequently, the average speed of tongue body during the transition was significantly smaller in the geminate context irrespective of place and manner of consonant articulation.

The main focus of the present work is on the coordination between the two coproduced articulatory movements, a bilabial gesture and a (fully or partially) coproduced lingual gesture realizing the transition between the flanking vowels. To date, only a handful of studies have investigated the influence of consonantal quantity on temporal aspects of this coordination. A majority of articulatory studies have confirmed the classic observation of Öhman (1966) that the inter-vocalic transition usually starts around (usually after) the onset of the consonantal gesture but, as a rule, before the attainment of consonantal target (e.g., a bilabial constriction) for the medial consonant. Looking at the temporal magnitude of this phenomenon, Gili Fivela *et al*. (2007) report that for Italian the vowel transition from /i/ to /a/ starts later relative to the (acoustic) onset of bilabial *constriction* for geminates compared to. Also, in their data *the interval from the onset of the bilabial gesture to the onset of vowel transition* (called V2LAG in

the present work, see also Figure 10) is longer[1] for geminates than for singletons (significantly so only for the normal rate).

The same phenomena were investigated by Löfqvist (2006) in his work on Japanese short and long nasal bilabials in /kV(m)mV/ sequences including five speakers and three vocalic contexts (/a-i/, /a-e/ and /e-a/). His measurements reveal a pattern which is not as clear. The duration of the interval from the onset of the tongue body movement towards the second vowel to the beginning of the acoustically determined oral closure does not show a straightforward dependence on consonant quantity; the trends are highly speaker dependent. Only slightly more consistent picture emerges in the case of the temporal coordination between the onsets of bilabial and vocalic gestures, our V2LAG measure. The inter-vocalic transition starts significantly later (greater V2LAG) for geminates in 7 out of 15 cases (3 vocalic contexts times 5 speakers) and significantly earlier in 2 cases. In 5 out of the remaining 6 non-significant cases, the mean V2LAG is greater for geminates. Finally, in a pilot study to this work (O'Dell et al., 2011a,b) we analyzed articulatory recordings of a single Finnish speaker (not included in the present study) uttering repetitions of sequences /ta(p)pi/ and /ti(p)pa/ and found significantly greater values of V2LAG for geminates compared to singletons.

In this paper we primarily investigate the dependence of inter-gestural temporal coordination on consonantal quantity in V(C)CV sequences, but, crucially, also on articulatory properties of the flanking vowels as well as on the consonant preceding the sequence. The articulatory recording and subsequent data analysis presented in Section 3 of this paper is motivated by the results of the dynamical modeling of gemination described in Section 2.

The modeling results come from the optimization based Embodied Task Dynamics (ETD) platform of Šimko & Cummins (2009, 2010), which combines the task dynamical implementation of Articulatory Phonology (Browman & Goldstein, 1992; Saltzman & Munhall, 1989) with principles of Lindblom's H&H theory and his concept of emergent phonology (Lindblom, 1990, 1999). In the ETD model, the temporal patterns of inter-gestural coordination emerge as solutions of an optimization task combining joint requirements to minimize articulatory effort and duration while maximizing clarity of the speech output. Rather than imposed in the form of, for example, phonologically motivated rules, inter-gestural phasing relations emerge from the interplay between functional requirements of communication and physical and physiological properties of the embodied vocal tract. The optimization task is implemented in the form of a cost function capturing the trade-offs between the production and perception oriented cost components in a flexible, adjustable manner. The model successfully accounts for several inter-gestural timing phenomena such as the dependency of phasing patterns in VCV sequences on the articulatory nature of the vocalic gestures and on lip-jaw-tongue synergies arising from anatomical links between the embodied articulators (Šimko & Cummins, 2009, 2010). In optimal gestural VCV sequences identified by the model, the relative lag of the onset of movement towards the second vowel with respect to the onset of consonantal gesture, i.e. the coordination measure V2LAG above, depends on the vocalic context: the tongue movement starts relatively later in /aCi/ sequences compared to /iCa/. This context-dependency faithfully captures corresponding patterns identified in articulatory recordings (Löfqvist & Gracco, 1999; O'Dell *et al.*, 2011a,b) suggesting that articulatory efficiency and perceptual efficacy play an important role in shaping coordination patterns in speech.

In the present work, we further investigate the ability of the ETD platform to account for context-dependency of inter-gestural timing, namely for phonological quantity contrast realized in a varied context in terms of articulatory nature of surrounding consonants. We explore (and,

[1] Note that the "duration" of this interval, V2LAG, as conceptualized here can be negative if the vocalic gesture starts before the consonantal one. The concept of "longer" is thus here extended to such negative "durations" in a natural way.

to a certain extent, explain) the emergent variations in gestural kinematics and phasing relations arising from articulatory synergies among gestures performed by an embodied vocal tract.

Consonantal quantity is modeled within the ETD platform through local adjustments of a parameter quantifying the trade-offs between the production and perception oriented cost components. Namely, increasing the premium placed on relative perceptual clarity of the consonant (lessening the chance of listener's skipping the segment when parsing the sequence) results in predictable lengthening of the consonantal closure in optimal sequences. Surprisingly, a continuous increase of the local premium leads to an *abrupt* change of the optimal coordination pattern. This discontinuity is manifested by a sudden increase of the closure duration for the consonant accompanied by a reorganization of inter-gestural phasing represented by the measure V2LAG. The qualitative difference between the two emerging patterns corresponds to the discrete phonological contrast between singletons and geminates. Moreover, the inter-gestural phasing relations in the optimal CVCV sequences are highly context-dependent. They show a strong dependency on the articulatory features of the vowels as well as of the sequence-initial consonant.

Our modeling results contribute to the ongoing debate on the so-called "phonetics-phonology problem" about the relationship between phonological (qualitative, discrete) and phonetic (variable, continuous) aspects of speech (Beckman & Kingston, 1990). We will show that under an explicitly formulated constraint of efficiency a discrete phonological contrast can emerge from continuous adjustments of an intentional parameter depicting local prominence. The contrast is instantiated as two distinct minima of the composite cost function, i.e., an affordance of the speech production and perception system to realize the vowel-consonant coarticulation in two discretely different ways.

In the following section we introduce a description of relevant aspects of the ETD model and present its predictions regarding the singleton/geminate contrast. Subsequently, we provide results of statistical analysis of articulatory recordings of Finnish CV(C)CV sequences and evaluate them in the light of the model predictions.

2.0 Model description and predictions

2.1 Embodied Task Dynamics model

The Embodied Task Dynamic model identifies articulatory patterns that are optimal with respect to competing demands of articulatory efficiency, perceptual clarity and overall durational requirements. The ETD platform is not conceptualized as a model of online speech production as, for example, the task dynamical implementation of Articulatory Phonology (Saltzman & Munhall, 1989; Nam et al., 2004) or the neural model of speech production DIVA (Guenther, 2001). Rather, it is intended to provide a computational platform for investigating qualitative aspects of inter- gestural coordination that result from perception and production constraints imposed by the embodied nature of the communicative system.

The optimal patterns are presented in terms of *gestural scores*: constellations of primitive articulatory movements, gestures, assembled in order to produce a given utterance. Each gestural score fully characterizes a given realization of an utterance. Formally, it consists of onset and offset activation times of constituent gestures and a relevant set of gestural dynamical parameters such as stiffness of the tract variable mass-spring system and corresponding gestural targets.

In the modeling platform discussed here, the gestural score is realized using a highly simplified model of the vocal tract schematically depicted in Figure 1. The model contains several articulators represented by masses: the upper and lower lips, the jaw, and two tongue components referred to as the tongue body and the tongue tip. The model captures only vertical movement of the model articulators; however, the sensitivity of various components to the movement of other articulators realistically reflects the horizontal arrangement of the vocal

tract. For example, the more frontal articulators, the tongue tip and the lower lip, are rendered more responsive to the vertical component of the jaw opening movement than the tongue body placed closer to the joint.

Note to Publisher: Insert Figure 1 about here

The model articulators are joined by critically damped springs representing muscles of the vocal tract (springs in Figure 1). These "muscle" connections reflect the anatomical linkages between the corresponding articulators of the human vocal tract: the lower lip and the tongue body are linked with the jaw, the tongue body in addition to the jaw also with the tongue tip. The upper lip is independently attached to the same frame of reference (the "skull") as the jaw. This arrangement captures basic degrees of freedom of the vocal tract. In fact, there are more degrees of freedom than necessarily required for any of the modeled speech gestures; the articulatory system is redundant. This redundancy allows coproduction of several speech gestures: the lip closure movement, for example, can be performed alongside the vocalic lingual articulation.

The gestural dynamics is implemented in a manner closely related to task dynamical implementation of Articulatory Phonology (Browman & Goldstein, 1992). Gestural targets are expressed as values of tract variables that capture linguistically meaningful spatial constellations of speech articulators. The *lip aperture* tract variable, for example, captures the distance between the lips, regardless of the positions of other vocal tract components. Bilabial closure is achieved when the value of this variable approaches zero. A bilabial gesture, i.e., the vocal tract *transition* towards the gestural target, is modeled by imposing critically damped mass-spring dynamics on the lip aperture tract variable with the target parameter set to a value less than 0 (-2 in this case). Trajectories of the individual vocal tract articulators are computed using a customized version of pseudo-inversion of the redundant mapping projecting articulatory position to the values of the tract variables (Šimko & Cummins, 2010).

The dynamics of the presented model differs from the traditional task dynamical implementation in several important ways (Šimko & Cummins, 2010). First, the individual articulators have assigned masses, approximately matching the masses of corresponding components of the human vocal tract. The dynamics of their movement thus reflects the physiological properties of the embodied speech system. Heavier articulators, like the jaw, move more slowly and with greater momentum than the lighter ones like the lips. The force required to move them is also greater than for the lighter articulators. Importantly, the force associated with the movement of all components can be evaluated by the system.

Second, the collisions between the articulators, and between them and the oral cavity boundaries are modeled through mutually repulsive forces, dynamically implemented in terms of a damping component inversely proportional to distance. Consequently, the articulators behave like elastic objects. When the model lips approach each other, the repulsion gradually slows down their movement until it balances the force induced by the bilabial closure. The complete closure is achieved when the repulsive force reaches a relatively small threshold, i.e., starts exerting a measurable influence on the lip dynamics. Any subsequent closing movement of the lips, significantly hampered by the repulsion force, is ascribed to elasticity of the lips (Löfqvist, 1996).

Third, each articulator is pulled towards its "speech-ready" position all the time. The articulator-specific speech-ready positions correspond to an average constellation with regard to the entire set of mastered gestures rather than a resting configuration of the vocal tract (attained for instance during quiet breathing). The pull towards the speech-ready state is modeled by critically damped mass-spring dynamics with targets set to appropriate positions for each individual articulator and a stiffness parameter considerably lower that that of any active gesture. As this dynamics is always on, the active gestures must overcome this weak re-setting force typically acting in the opposite direction. Consequently, the equilibrium points for an

articulator under the influence of a speech gesture are slightly offset compared to their dynamical targets. When no gesture acts on an articulator it slowly returns to the speech-ready position.

[READ TO HERE]

To guarantee that the active gestures exert an intended influence, an *overall stiffness* parameter is an "adjustable" parameter of the system (and subject to subsequent optimization, see below). For parsimony's sake, both the gestural and speech-ready stiffness parameters are fixed relative to the value of this single parameter. The speech-ready dynamics stiffness coefficients as well as the stiffness parameters for active gestural dynamics are thus defined as multiples of the overall stiffness value. The value of the multiplication (relative stiffness) coefficients has been decided in advance and is not adjusted during the optimization process. As a consequence, the ratios between dynamical stiffness of each pair of gestures in the sequence remains fixed; we shall return to possible consequences of this modeling decisions in Discussion.

The simplified anatomy of the model vocal tract realistically allows for implementing only a limited set of gestures representing linguistic contrasts. Each gesture is associated with an appropriate tract variable, its dynamical target, and the coefficients defining the relation of its dynamical stiffness to the overall stiffness parameter. Alongside the relative stiffness coefficients discussed above, the gestural targets are also treated as fixed (e.g., learned during speech acquisition) and not as part of gestural scores that are subject to optimization. In addition to gestural activation onset and offset times, the overall stiffness is thus a single optimized parameter.

The model in the presented form distinguishes three tract variables: lip aperture, tongue tip position (relative to the alveolar ridge) and vertical position of the tongue body. Using these tract variables, the model depicts four linguistically meaningful gestures: a bilabial closure (a convergence of lip aperture towards the gestural target 0, i.e., collision of the lips), alveolar closure (the tongue tip tract variable converging towards its target associated with the oral cavity boundary) and two vocalic gestures associated with high and low targets of the tongue body tract variable, respectively. These four articulatory gestures are labeled in this work as /p/, /t/, /i/, and /a/, respectively. Voicing and nasalization are not implemented in the model.

Each active gesture imposes critically damped mass-spring dynamics on the associated tract variable. Unlike the traditional task dynamic implementation of articulatory phonology (Browman & Goldstein, 1992), the behavior of model articulators is not determined solely by gestural dynamical parameters, i.e., gestural target and stiffness coefficient. The behavior is affected by always-active speech ready dynamics and, importantly, it also reflects masses associated with each model articulator. The masses act as additional parameters of critically damped mass-spring dynamics of individual articulators linked to the tract variable dynamics through a pseudo-inversion of the articulator-to-tract-variable mapping (cf. Saltzman & Munhall, 1989). In the embodied task dynamical system it is thus possible to evaluate forces acting on individual model articulators during realization of a given gestural score (as a product of mass and acceleration dynamically imposed on the articulators). For the full details of the model vocal tract anatomy and definition of its dynamics, see Šimko (2009) and Šimko & Cummins (2010).

Figure 2 shows a gestural score for a sequence /tapi/ and corresponding actions of model articulators as evaluated by the embodied task dynamics. While realizing the given sequence, the score and the corresponding articulator trajectories have not been optimized; it is suboptimal with respect to the criteria defined below. Comparison with an optimal score for the same sequence shown in Figure 8 reveals the effects of the optimization process described in the following section.

Note to Publisher: Insert Figure 2 about here

2.2 Composite cost function and optimization

A gestural score captures onset and offsets of all gestures participating in production of the given utterance and the overall stiffness parameter. Alongside the remaining parameters of the model, which remain fixed, it fully determines the kinematics of the model vocal tract articulators. The objective of the optimization process is to find the optimal gestural score that simultaneously minimizes articulatory effort, effort associated with parsing of the resulting utterance and the overall duration of the utterance.

Each gestural score is assigned a cost combining measures of articulatory and parsing effort and duration. The compound cost C associated with the gestural score is a linear combination of three cost components

$$C = \alpha_E E + \alpha_P P + \alpha_D D, \quad (1)$$

where E is the production articulatory cost of realizing the gestural score, P is the parsing cost, and D is the overall duration of the utterance represented by the gestural score.

The production cost E associated with articulatory effort is evaluated as overall force expenditure during the realization of the given gestural score. As described above, the embodied character of the task dynamical model used here facilitates evaluating realistic forces acting on each articulator represented by a mass. The time-integral of the magnitude of this force over the duration of the utterance captures the overall force exerted on this articulator. The cost E is defined as a sum of these overall force measures for all model articulators.

As it has been repeatedly pointed out by speech researchers, establishing the role of articulatory effort and its conservation – as assumed here – is far from straightforward (see, e.g., Pouplier (2012) for a lucid discussion of this issue). There are at least two aspects to this. First, it is not clear that production efficiency has any influence of shaping speech production patterns. After all, unlike more robust types of physiological action like running or lifting objects, speech articulation is effected by comparatively small forces exerted by musculature specialized for fatigue resistance. Attempts to directly evaluate metabolic cost associated with speech have up to date lead to inconclusive results at best (Moon & Lindblom, 2003). An accessible way to evaluate the influence of efficiency constraints posed by embodied articulation is a modeling approach such as presented in this work. Although an agreement of details of inter-gestural sequencing in human subjects and predictions of the model incorporating such constraints does not constitute a proof, it contributes to a growing body of evidence interpreting speech as a task oriented adaptive action subject to a combination of efficiency requirements, including conservation of articulatory effort.

The second issue directly impacts modeling decisions. Some intuitive measures of articulatory effort, like a distance traveled by an end effector in order to achieve sufficient contrast, do not take into account complex biomechanical properties of the vocal tract. For example, Perrier *et al*. (2003) have shown that a looping motion of the tongue body during the production of an [ugu] sequence is efficient with respect to biomechanical factors although it clearly does not constitute the shortest possible trajectory realizing the sequence. Our definition of the cost component E reflects this insight: the effort is evaluated in terms of forces exerted on embodied model articulators rather than in terms of kinematic properties of end effector trajectory.

Finally, it is important to note that the requirement of articulation parsimony in counterbalanced in a context dependent way by the effects of two additional cost components.

The duration cost D is defined simply as the overall duration of the realization of the gestural score in seconds. The parsing cost component P is a measure of clarity and is designed to reflect the effort exerted by the listener in order to recognize the intended sequence. Its precise

definition is closely connected with modeling the singleton/geminate contrast, therefore we describe it in more detail below. To preview, the cost increases with very short and imprecise articulation of gestures.

Equation (1) is a quantitative expression of the basic idea of the Lindblom's (1990) H&H Theory. Phonetic variation is a consequence of tradeoffs between conflicting requirements of production efficiency (captured by *E*) and perceptual clarity (*P*). In addition, the durational component *D* incorporates demands on a shared resource – time – and acts as a partly independent means of eliciting variation in speaking rate dimension (for more thorough analysis of influence of each component, see Šimko & Cummins, 2011). The coefficients α_E, α_P, and α_D assign weights to individual aspects and express intentional control of variation along hypo- and hyper-articulation scale as well as speaking rate. For example, increasing the parsing cost weight α_P puts greater weight on clarity: the utterance should be easier for the listener to parse. Consequently, the optimal gestural score will prescribe longer, more precisely articulated speech segments. Similarly, increasing α_D will favor gestural scores with shorter realizations resulting in faster speech. Importantly, the variations can be elicited not only by global but also by *local* weight adjustments. Beňuš & Šimko (this issue) used local lowering of α_D to elicit articulatory patterns reproducing those accompanying prosodic breaks. In this work we explore how the local adjustments of α_p can account for phenomena associated with quantity contrast.

The optimization procedure for finding the optimal gestural score uses an adapted simulated-annealing process with the overall cost *C* as its objective function. Starting with an arbitrary gestural score, sub-optimally realizing the required sequence of gestures, the procedure searches in a gestural score's vicinity for a solution carrying smaller cost than the previous candidate. This process continues iteratively until no further progress can be made. At each step, the candidate gestural score (i.e., the gestures' onset and offset times plus the overall stiffness) is randomly perturbed. This helps the process avoid "getting stuck" in locally optimal solutions. The gestural score that cannot be further improved provides the globally optimal solution minimizing the overall cost function.

2.3 Parsing cost component and consonantal quantity contrast

The parsing cost component is an estimate of an effort required by a listener to successfully parse the intended utterance. As the model discussed here is purely articulatory, this effort is estimated through positional and durational characteristics of realized gestures. This approach abstracts away from the well-known nonlinearities in articulation-to-acoustics mapping and in auditory processing (Stevens, 1989). We assume that the parsing effort is inversely proportional to the *duration* of each individual speech segment and, for vowels, also (linearly) proportional to the extent of *undershoot*, i.e. the minimal distance of the relevant articulator from the given vocalic articulatory target during the gesture's realization interval.

The *realization interval* of a gesture is a stretch of time during which the intended consequences of the gesture are achieved (realized). The realization interval thus typically lags in time behind the gesture's activation interval prescribed in gestural score (see Figure 2). Conceptually, realization interval is related to a phone segment as traditionally delimited in speech signal by phoneticians.

In the formal definition of realization interval, the purely articulatory character of the model enforces several simplifying departures from the traditional conceptualization of phonetic segment. For stop consonants, the realization interval is defined as a period during which the vocal tract closure is achieved (taking the closure itself as the intended realization of the bilabial or alveolar gesture and ignoring the significance of the following burst). Technically, this period is delimited using the repulsive force implemented as a damping component of gestural dynamics and acting on model articulators when they approach each other or the oral cavity boundaries. The closure is attained when the repulsive force exceeds a small threshold beyond which the tissue elasticity impacts the articulator behavior. Vocalic gestures, on the other hand,

are considered as realized while not occluded by a consonantal realization interval *and* when the relevant end effector (tongue body) is more than two-thirds of the way between its speech-ready position and the associated gestural target.

In the model, duration of the realization interval directly impacts the value of the parsing cost component. Extremely short realization interval dramatically increases the likelihood that the listener misinterprets the intended consequences of the underlying gesture, and is thus associated with a very high cost. As the realization interval gets longer, the impact of duration on perceiving the given gesture gradually decreases. The model accounts for this effect through a function $P(g)$ capturing the impact of the perceived duration of gesture g on the parsing cost defined as

$$P(g) = 0.001/T_g,$$

where T_g is the duration of realization interval of the gesture (see Figure 3)[2].

For a stop gesture g, $P(g)$ fully encapsulates its contribution to the effort of parsing the utterance. In effect, we assume that longer closures make stops relatively more prominent (in a non-linear fashion) and thus facilitate identification of their presence in the sequence. As the model does not account for differences in manner and the two modeled consonantal gestures are produced with different end effectors – bringing along appropriate spectral cues in the form of formant transitions – closure duration is seen as sufficient cue for distinguishing between "poorly" and "well" articulated consonants. This simplifying assumption is supported by the well-established finding that the likelihood of listeners recognizing a stop consonant in a sequence increases non-linearly with the duration of silence. For example, Best *et al.* (1981) reported that even with the appropriate spectral cues the probability of interpreting the sequence /s/ + silence + /ei/ as /stei/ by listeners increases non-linearly with the silence interval duration increasing from 0 to approximately 70 ms. In other words, very short durations of the closure make the task of correctly parsing the sequence progressively difficult.

For vowels, produced by the same model end effector (tongue body), the listener's task to successfully parse the segment is not limited to detecting its presence; they also have to correctly judge the vowel quality. Consequently, the parsing cost component is for the vowels complemented by a measure of articulatory precision achieved during the realization interval (the details can be found in, e.g., Šimko & Cummins, 2010).

Note to Publisher: Insert Figure 3 about here

The parsing cost associated with the utterance depicted by a gestural score is defined as a linear combination of the contributions of all the individual gestures g:

$$P = \Sigma g\, \beta_g P(g).$$

The coefficients β_g depict *local premiums* on the parsing cost and they weight relative contributions of gestures. First, they are used to compensate for an influence of precision estimate for vowels (as a consequence of its presence, the vocalic contributions $P(g)$ are considerably smaller than consonantal ones) and to adjust relative durations of vowels and consonants in an optimal realization of an utterance.

[2] This definition of the durational component of parsing cost component differs from the models presented in our previous work (e.g., Šimko, 2009; Šimko & Cummins, 2010, 2011; O'Dell *etal.*, 2011a,b) where we used a different function (arc tangent) with similar mathematical properties to the reciprocal introduced here. While the choice of the particular function has consequences for interpretation of emergence of singleton-geminate contrast in terms of local prominence, the predictions and kinematic characteristics presented in this section are qualitatively the same also for the older version of the model.

More importantly, the weights influence *relative prominence* of gestures. Increasing the local premium means placing extra incentive on parseability of the associated gesture and results in rendering the gesture relatively more prominent in the optimal realization of the given utterance. This feature of the model plays a crucial role in the context of the presented study.

Recall that for stop consonants the parsing cost contribution rests solely with the duration of the consonantal closure. Increasing the local premium β_g for a selected consonantal gesture will – in the optimal gestural score – inevitably bring about longer closure duration. Increasing the local premium is equivalent to decreasing the slope of the non-linear relationship between duration of a gesture's realization interval and the associated cost. Figure 3 illustrates this fact. When the value of β_g is increased for the consonant (e.g., from 2 to 15 as suggested in the figure), the parsing cost for the previously optimal consonant duration (marked as "singleton") becomes prohibitively high relative to the parsing cost of surrounding gestures for which no change in local premium has been effected. To achieve an optimal equilibrium, the duration of the consonantal gesture must increase. In fact, as the increase is inevitably associated with an extra effort and duration cost, the optimization process finds an optimal *longer* duration – and, as we shall see, also adjusts the phasing relations with the surrounding gestures – for the consonantal gesture.

The lengthening of the closure straightforwardly corresponds to the primary durational characteristic of contrast between singletons and geminates. We have deployed this insight as the central tool for modeling the phonological contrast. In the following section we present the modeling results and summarize emergent phonetic characteristics of the singleton-geminate contrast as captured by the model.

2.4 Emergent bimodality: Results of modeling the contrast between singletons and geminates
Figure 4 shows the optimal gestural scores and corresponding articulatory trajectories for a V_1CV_2 sequence /api/. The difference between the gestural scores result from the differences between the local premiums placed on parsing cost for the bilabial gesture /p/. For the left hand side score its value was set to 2 ("singleton" value), for the right hand "geminate" score side to 15 (the weight for vocalic gestures is set to 15 in both cases). All remaining parameters of the gestural dynamics and optimization process ($\alpha_E = \alpha_P = 1$, $\alpha_D = 3$) are identical for both cases.

Note to Publisher: Insert Figure 4 about here

As expected, the resultant optimal duration of closure – marked by vertical lines, onset by the full and offset by the dashed one, respectively – is greater for the "geminate" setting (118 ms) than for the "singleton" one (54 ms). This is achieved by a later offset of the "geminate" bilabial gesture /p/ after the closure achievement compared to the "singleton" one. This phenomenon is a straightforward and intended consequence of our modeling approach.

More interestingly, the optimal scores also differ in another characteristic of temporal sequencing: the lag of V_2-gesture (/i/) activation onset with respect to the onset of the consonantal /p/-gesture is considerably greater in the "geminate" case compared to the "singleton" case. As we refer to this temporal coordination measure throughout the remaining text, we shall introduce here a convenient abbreviation applied both to gestural scores produced by the model and to measurements of articulatory recordings:

V2LAG is the duration of the interval from activation onset of the bilabial gesture under consideration to the onset of activation of the vocalic gesture immediately following the bilabial; in formal terms, time of vocalic gesture onset minus time of bilabial gesture onset.

Please note that the V2LAG measure is (somewhat counter-intuitively) negative in cases where the onset of the vocalic gesture *precedes* that of the bilabial one.

In the optimal scores depicted in **Virhe. Viitteen lähdettä ei löytynyt.**, V2LAG equals 34 ms for the "singleton" and 67 ms for the "geminate". Similarly, for the /ipa/ sequences in Figure 5, V2LAG is greater for the "geminate" case (right pane) (38 ms) than for the "singleton" one (14 ms). The closure durations in the optimal constellations in **Virhe. Viitteen lähdettä ei löytynyt.** are 46 ms for the "singleton" and 100 ms for the "geminate".

Please note, that for both pairs of "singleton" and "geminate" optimal sequences, V2LAG is greater for /api/ than for /ipa/. This dependency of inter-gestural timing on vocalic articulatory context accounted for by the model and matching articulatory measurements (Löfqvist & Gracco, 1999; O'Dell et al., 2011a,b), has been described by Šimko & Cummins (2010, 2011). We shall investigate this context sensitivity along with sensitivity of coordination patterns to even wider segmental context (preceding consonant) in the following section.

Note to Publisher: Insert Figure 5 about here

Returning to the dependency of V2LAG on consonantal quantity, this property of inter-gestural timing in optimal gestural scores is open to a straightforward, common-sense interpretation: a sequence medial geminate means longer bilabial closure, therefore the intervocalic transition has time to start relatively later than when the consonant is a shorter singleton. This analysis assumes a direct relationship between duration of closure (or bilabial gesture) and our measure of inter-gestural coordination V2LAG.

The model determines the duration of the closure by the local premium placed on perceptual prominence of the bilabial consonant. As mentioned above, the optimal scores depicting the singleton-geminate contrast use values of 2 and 15 for this parameter. Unlike articulatory recordings (human speakers as a rule cannot produce an exact, finely grained continuous scale of singleton-geminate consonants), the modeling paradigm allows us to visualize the relationship between the closure duration and inter-gestural coordination even for intermediate and extreme situations.

Solid circles in **Virhe. Viitteen lähdettä ei löytynyt.**A show the values of V2LAG in *optimal* gestural sequences /api/ computed for the local premium ranging from 1 to 16 (step 0.5). For each value of the premium, we identified the optimal gestural score using the procedure described above, and extracted the value V2LAG (as the lag of the onset of /i/-gesture after the onset of /p/-gesture in the optimal score). As seen from the plot, V2LAG increases with the local premium – the more prominence placed on the bilabial, the relatively later the intervocalic transition starts. The relationship, however, is far from linear. Up to premium value of 3.5 the lag slowly increases from about 32 to 47 ms. Then, a small increase of the premium is accompanied by a sudden "jump" of the optimal V2LAG value to approximately 64 ms, afterwards the V2LAG value remains in the area around 65 ms despite the increasing premium. An area with no optimal values of V2LAG is marked in the figure as a "gap".

In fact, for the local premium values between 2.5 and 4 the optimal gestural score is not the only stable solution of the cost minimization task. Empty gray circles mark additional constellations that are locally optimal. (These solutions were calculated for the premium values between 2.5 and 4 by limiting the range of the optimization procedure.) The premium values of 2.5 and 4 thus delimit a *bistable region* of control variable depicted in Figure 6A. Within this region, the overall cost function has two minima: a global minimum depicted by the full circle and an additional (local) minimum shown by the empty gray circle. In other words, as the prominence of the bilabial gesture increases, the single solution of the optimization task bifurcates to two stable, locally optimal constellations. Subsequently, one of these stable patterns disappears, once again leaving a single solution, qualitatively different from the original one.[3]

[3] This phenomenon is reminiscent of the concept of reorganization of 'attractor landscape' discussed for example in Gafos' (2006) treatment of phonological and phonetic characteristics of final devoicing

Note to Publisher: Insert Figure 6 about here

The difference between these two constellations is shown in **Virhe. Viitteen lähdettä ei löytynyt.**B depicting the relationship between V2LAG and the duration of bilabial closure CLDUR extracted from the same (locally) optimal scores used in **Virhe. Viitteen lähdettä ei löytynyt.**A (the corresponding premium values used to find the sequences are shown next to the data points). The "jump" in V2LAG is accompanied by a similar shift in the closure durations that abruptly increase by approximately 20 ms. This is qualitatively analogous to the distinction between singleton and geminate stops. Therefore, we designate as "singletons" the patterns arising from a premium value less than 4 (enclosed by a dashed ellipse in Figure 6B) and as "geminates" those determined by greater values (solid ellipse). The gestural scores plotted in **Virhe. Viitteen lähdettä ei löytynyt.** are the optimal constellations from which the V2LAG and CLDUR values marked as 2 and 15, respectively, were extracted.

Note to Publisher: Insert Figure 7 about here

The /ipa/ sequences display qualitatively similar behavior, albeit slightly differing in details. As shown in **Virhe. Viitteen lähdettä ei löytynyt.**, the qualitative jump or bifurcation, occurs for slightly greater values of the local premium (bistable region lies between 5.5 and 6.5), and is much less prominent. Nevertheless, the values of the inter-gestural coordination measure, V2LAG, are relatively constant for small (2—5.5) and large (over 8) values of the premium. Varying the local premium for the bilabial gesture once again leads to two stable inter-gestural coordination patterns. The situation is analogous to the postulated quantal nature of articulatory-to-acoustic mapping (Stevens, 1989).

Virhe. Viitteen lähdettä ei löytynyt.B reveals that the surface characteristics of these two stable regions again correspond to the difference marking the consonantal quantity contrast: the closure durations in the optimal gestural scores obtained using small premium values are considerably shorter than in those enforcing higher prominence of the bilabial gesture.

The simple requirement of increased perceptual prominence of geminates compared to singletons leads to more than a mere quantitative difference in closure duration between these two phonological categories. The optimization modeling results suggest a qualitative difference, a genuine bimodality, in inter-gestural coordination patterns. The distinction is marked by a greater lag of intervocalic transition relative to the onset of bilabial gesture in geminates than singletons (V2LAG). This hallmark of quantity emerging in our modeling results serves as one of the principal predictions to be verified using human articulatory data.

2.5 Consonantal context: Dependence of sequencing on preceding consonant

The results outlined above suggest dependence of inter-gestural timing in VCV sequences with bilabial stop on the vocalic context (/a-i/ vs. /i-a/) and on the prominence level ascribed to the stop consonant. In this section we explore sensitivity of the sequencing pattern to a wider gestural context, namely to the consonant immediately preceding the VCV sequence.

This inquiry is motivated by the results of articulatory analysis of a Slovak corpus reported by Šimko *et al*. (2011) suggesting an influence of gesture-initial lip opening on temporal sequencing details. Presence of another bilabial stop just before a VCV sequence containing a bilabial consonant could lead to a decrease of the lip opening during the first vowel of the

---

interpreted in terms of non-linear dynamics. Plotting overall cost C as a function of duration of bilabial gesture in the bistable region would give a potential well with two local minima, qualitatively equivalent to Gafos's potential well model for voiced vs. voiceless stops. The discovery of bistability in the physically realistic model provides a grounding for the dynamic systems account proposed by Gafos and others.

sequence and subsequently impact the inter-gestural timing. Moreover, a sequence of multiple bilabial closures could lead to a phenomenon of "crowding" discussed (along with its consequences for inter-gestural phasing) by Beňuš & Šimko (this issue).

Note to Publisher: Insert Figure 8 about here

Note to Publisher: Insert Figure 9 about here

Figures 8 and 9 show the optimal scores for sequences using /api/ and /ipa/ with the same "singleton" and "geminate" settings used in Figures 4 and 5 each preceded by a /t/ or /p/ gesture. Table 1 offers kinematic characteristics of the lip closing gesture for C_2 (/p/ or /pp/) and for vocalic transition V_2 as well as durations of the "acoustic" interval for V_1 (from the offset of C_1-closure to the beginning of C_2-closure) and duration of the C_2-closure. The kinematic characteristics were computed in a way fully compatible with the measurement of empirical articulatory data analyzed in the following section (see Section 3.2 for details). The durational measures based on onsets and offsets of relevant movements reported in Table 1 are thus slightly different than the gestural onsets and offsets suggested by the gestural score. Due to inertia of articulators, the lip closure movement, for example, starts as a rule slightly later than the dynamic influence driving the lips together. For visual analysis, however, the optimal gestural scores are sufficient.

Table 1. *Kinematic and "acoustic" characteristics of the simulated sequences shown in Figures 8 and 9.*

	"acoustic measures"		consonant gesture			V_2 gesture			
	closure duration	V1 duration	duration	displ.	p. vel.	duration	displ.	p.vel.	V2LAG
pipa	55	69	45	3.0	97	168	13.7	149	-6
pippa	107	72	84	4.1	86	182	13.6	135	-7
papi	66	92	74	3.8	72	177	13.5	135	-2
pappi	128	68	103	3.5	57	191	13.5	126	13
tipa	53	75	67	6.7	167	169	13.7	148	12
tippa	107	73	110	8.0	147	183	13.6	137	20
tapi	69	97	109	9.4	170	181	13.7	135	34
tappi	129	75	147	10.2	158	194	13.6	125	58

The figures and measures in Table 1 offer several observations. Most characteristics show a strong dependence on the three influences under investigation here: vocalic and consonantal articulatory context, and gemination. In our discussion of the qualitative nature of these influences below, we compare all sequences differing solely in the given aspect. To evaluate the effect of vocalic context, for example, we thus compare the relevant measures for pairs /tapi/-/tipa/, /tappi/-/tippa/, /papi/-/pipa/ and /pappi/-/pippa/.

The following evaluation of the modeling results can be seen as a list of predictions of our model. We will therefore number the following paragraphs so that we can refer to particular predictions in the following section dedicated to data analysis.

(1) The first column in Table 1 shows that closure duration is considerably influenced by gemination as introduced in the model. The ratios between geminate / singleton constriction duration vary between 1.86 for /tapi/-/tappi/ pair to 2.0 for /tipa/-/tippa/. (The particular values of the local premium parameter, 2 for singletons and 15 for geminates, were in fact selected to achieve the ratios of approximately this magnitude. Therefore, the ratio values cannot be interpreted as emergent results in the same sense as most of the observations below.). There is no clear context dependent pattern, although the ratios are somewhat larger in the /i-a/ context.

The constrictions themselves are longer for /a-i/ than for /i-a/, and to a very small extent in sequences beginning with /t/ compared to /p/-sequences.

(2) In the /i-a/ context, the durations of V_1 are only minimally influenced by gemination (somewhat longer /i/ in /pippa/ than in /pipa/ and tiny bit shorter in /tippa/ than in /tipa/). In the opposite /a-i/ context, however, gemination shortens the preceding vowel considerably. With the exception of the /pippa/-/pappi/ pair, /a/ is longer than /i/ in sequences differing solely in vocalic context. V_1 duration is also influenced by consonantal context, in comparable sentences V_1 is longer when it is preceded by /t/ than by /p/.

(3) Turning our attention to properties of the consonantal articulatory gesture, the model predicts a consistent lengthening of its duration for long consonants. The lip closing movement is also, in fact to a larger degree, influenced by the sequence-initial consonant, being longer in the /t/-context than in the /p/-context. The vocalic context also plays a role, lip closing takes longer in /a-i/ than in /i-a/ sequences.

(4) The lip displacement during the closing gesture – the difference between maximal and minimal lip aperture at the onset and offset of the movement – is greater for geminates than for singletons (with the exception of /papi/-/pappi/), greater for /a-i/ than for /i-a/ (with the exception of /pippa/-/pappi/) and considerably greater for /t/ than for /p/.

(5) In the simulated sequences, the closing gesture reaches a higher peak velocity for singletons than geminates, and to even greater extent for the /t/ sequences than for the /p/ sequences. The effect of vocalic context shows an interesting pattern of interaction with the sequence initial consonant: in sequences staring with /p/, the peak velocity is greater for /i-a/ than for /a-i/, in /t/-sequences it is higher for /a-i/ than for /i-a/.

The next three paragraphs concern the kinematic measurements of the V_2 gesture, i.e., of the articulatory transition between the vowels flanking the consonant undergoing gemination.

(6) Compared to the kinematic measures of the consonant movement, the consonantal context (leading /p/ vs. /t/) has very little influence on the vowel transition. Duration of the V_2 gesture is influenced by gemination (longer for geminates than singletons) and the articulatory nature of the flanking vowels themselves (longer for /a-i/ than for /i-a/).

(7) The model does not predict any influence of gemination and surrounding articulatory context on the overall displacement of tongue body during the intervocalic transition (the distance between its most extreme positions during realization intervals of the flanking vowels). The small numerical differences are within the error margins of our stochastic optimization procedure.

(8) While consonantal context does not show any influence on the peak velocity of the transition, both gemination and vocalic context do. For all relevant pairs, the peak velocity is higher for singleton sequences than for geminate ones. Also, the tongue body reaches a higher peak velocity in the optimal sequences when moving from /i/ to /a/ than during the opposite movement.

(9) Finally, we look at predictions regarding context and gemination effects on the inter-gestural coordination measure V2LAG, central in the context of modeling the singleton-gemination contrast in this work. The relationship between the consonant-vowel coordination and gemination described in detail in Section 2.4 (V2LAG greater for geminates than singletons) holds also for VCV sequences preceded by a consonant, with the exception of /pipa/-/pippa/. For this pair, V2LAG is actually marginally smaller in the geminate context. Vowel context also continues to exert an effect, with V2LAG smaller for /i-a/ than for /a-i/. Consonantal context has, numerically, the greatest effect of all. While the sequences starting with /t/ show inter-gestural

phasing relations comparable to the VCV sequences presented earlier (see Figures 1, 2 and 5, 6), sequence-initial bilabial /p/ strongly affects the coordination. The second /p/ in these sequences is "pushed to the right" resulting in smaller V2LAG (i.e., smaller lead of lip closing before vowel movement).

The dependency of the relative phasing between the consonantal and vocalic gestures on quantity and articulatory nature of surrounding gestures (point 9 above) is the key prediction of the model. The first reason is that the model is explicitly conceived to account for temporal details of inter-gestural sequencing: the variables of the objective function are primarily the onsets and offsets of gestures' activations. The ability of the model to faithfully reproduce kinematic properties of individual gestures is compromised by decisions regarding gestural dynamics and its parameters: the model employs a simple critically damped mass-spring dynamics and optimization process cannot independently fine-tune the stiffness parameters of individual gestures (see Section 2.1). Secondly, as explained in Section 2.4 the two distinct organizational patterns for singletons and geminates are identified by the sudden "jump" in the value of V2LAG. Therefore, this measure serves as a benchmark for evaluating the possible plausibility of this finding.

Nevertheless, the agreement or disagreement of the remaining predictions (1-8) above with the empirical data will contribute to overall evaluation of the modeling decisions and might reveal shortcomings to be addressed in the future. For an overview of the correspondence between the predictions and data analysis, see Table 2.

3.0 Articulatory measurements

3.1 Linguistic material and recording procedure

The analyzed linguistic material consists of two-syllable words /$C_1V_1.C_2V_2$/, were the vowels V_1 and V_2 are /a/ and /i/, or /i/ and /a/, respectively, consonant C_1 is /p/ or /t/, and consonant C_2 is singleton or geminate voiceless bilabial /p/ or /pp/. All possible combinations thus yield 8 tokens – /tapi/, /tappi/, /tipa/, /tippa/, /papi/, /pappi/, /pipa/, /pippa/. Although all tokens are phonologically well formed in Finnish, only three are actually occurring word (*tappi* means 'a pin', *tippa* means 'a drop' and *pappi* means 'a priest'; *pipa* means 'wool cap' in some dialects, but not in the variety spoken by our speakers).

These tokens were recorded in two conditions:

1. as simple *repetitions* – speakers were simply instructed to repeat each two-syllable target word 10 times, and
2. embedded in a Finnish carrier sentence (two tokens in each sentence) *Hän sanoi ___ sekä ___*. 'S/he said ____ as well as ____.' In every sentence, a test word thus appears in *sentence medial* and *sentence final* position. To facilitate articulatory analysis, each test word in the carrier phrase was preceded by the (meaningless) word /kekV/, where V stands for /a/ or /i/, depending on the first vowel V_1 in the test word; the vowel V was always different from V_1. An example of the full sentence is "Hän sanoi keka tipa sekä keka tippa." To make the task somewhat easier, the two test words in the sentence differed only in consonant C_1 (/p/ and /t/) or consonant C_2 (singleton and geminate). To avoid a possible effect of read speech subjects were shown two target phrases (e.g., keka tipa – keka tippa) for 2 seconds and were instructed to wait until the phrases disappeared from the screen before uttering the given sentence. (We allowed subjects to start the sentence while the phrases were still on the screen, when the repetitive nature of the task posed too large a demand on their attention at later stages of the recording session.) Finally, the sentences were uttered at *normal* and self-imposed *fast* speaking rates.

Four subjects were recorded, two females (S1 and S3) and two males (S2 and S4. All are native speakers of Finnish from Helsinki or surrounding area (Uusimaa, Finland). None of them reported any speech or hearing disorders. Subject S2 is the third author.

The movement of the lips, the tongue and the jaw were recorded using electromagnetic articulography (EMA, Carstens AG500) at the University of Helsinki, Finland. The relevant receivers were placed on the midsagittal plane of the vocal tract; the lip sensors were placed above and below the vermilion border of the upper and lower lip, the jaw sensor below the lower incisors, and three tongue sensors at the tip of the tongue, at the rear portion of the tongue (as far back as confortable for the subject) and half way between these two points. In this analysis, we use the middle tongue sensor (referred to as tongue body, TB) and the upper and lower lip sensors (UL and LL, respectively). The articulatory movement signals, recorded at a sampling rate of 200 Hz, were converted to position values using Carstens' Calcpos program. Subsequently, the position signals were corrected using the Amplitude Adjustment algorithm of Hoole & Zierdt (2010); their program was also used for adjustment to head movement using additional sensors attached to the bridge of the subject's nose and behind each ear. The resulting postion signals were smoothed using an 8-point Bartlett window and subsequently up-sampled to 1000 Hz using cubic spline interpolation.

Due to technical difficulties during the recording sessions, we didn't manage to record and postprocess succesfully the same number of tokens for each subject. The numbers of tokens analyzed in this work for each subject and each condition range between 3 and 7 for each individual condition (such as rate—position combination) in sentence tokens and 6 to 21 tokens in repetitions.

3.2 Labeling

A single annotator identified the onset and offset of the bilabial closure for C_2 (/p/ or /pp/) and the closure (bilabial /p/ or alveolar /t/) for C_1 in all tokens, using the acoustic signal recorded synchronously with the articulatory data. The acoustic closure onsets and offsets were used as anchors for automatic articulatory labeling implemented in Matlab.

A lip aperture measure was calculated as the Euclidean distance between the positions of the lower lip and upper lip sensors. The onset of the lip closure movement for the bilabial stop /p/ or /pp/ was found as the velocity zero-crossing of the lip aperture signal, occurring before the acoustic closure. Offset of the closing movement was identified with the instant of maximal compression, i.e., minimal value of lip aperture during the acoustic closure. Duration of the closing movement was calculated as the duration of the interval from movement onset to movement offset. Displacement of lip aperture is the absolute value of the difference between lip aperture at the onset and the offset of the closing movement.

Analysis of tongue body movement was complicated by the presence of plateaux – intervals of minimal movement – of the TB sensor during the production of vowels preceding and following the stop C_2, mostly in the geminate context. Therefore, the onsets and offsets of the lingual movement were identified using tangential velocity of the sensor (cf. Löfqvist & Gracco, 1999). The onset of the movement was taken to be the last local minimum of tangential velocity (not necessarily zero) preceding a substantial movement of the articulator; the offset was marked in the corresponding fashion. The displacement of the tongue body during the movement is the Euclidean distance between TB sensor positions at the onset and offset of the transition movement.

Just as in the modeling context, the inter-gestural coordination measure V2LAG was computed as the onset time of the vowel transition movement minus the onset time of the bilabial closing movement.

Peak velocity of the lip closing gesture for C_2 was identified as the maximum velocity magnitude of the lip aperture signal between the onset and offset of the closing movement. Similarly, for the TB movement between the flanking vowels, the peak velocity was found as the maximum tangential velocity during this movement.

Note to Publisher: Insert Figure 10 about here

Virhe. Viitteen lähdettä ei löytynyt. illustrates the procedure using one of the articulatory recordings analyzed in this work.

3.3 Results

The main focus of our data analysis is on the acoustic and articulatory measures that are relevant for the predictions of our modeling effort. A more thorough analysis of the data with the aim to further contribute to our knowledge of Finnish geminate articulation will be presented in the near future.

We used analysis of variance (ANOVA) to assess the influence of three independent variables on the measures of interest: *gemination*, vocalic context (*VV-context*, /a-i/ vs. /i-a/) and consonantal context (*C1-context*, that is the place of articulation of the initial consonant in CV(C)CV sequences, i.e., /p/ vs. /t/). Reflecting the relatively low number of subjects and differences between elicitation methods, ANOVA was performed separately for sentences and for repetitions, and separately for each speaker. In order to help reader keep track of the rather large number of results, we limit the presentation to main effects, reporting significant interactions only when necessary. The analysis results are shown in more detail in Appendix.

For the tokens embedded in sentences we also included in the ANOVA model the influence of *tempo* (normal vs. fast speaking rate) and *position* of the sequence within the sentence (medial vs. final); for repetitions we included the token's *place in the series* (1—10). In order to make the following text less cumbersome, we do not report the effects of *tempo*, *position* and *place in the series* on the dependent variables of interest except for some interactions. In any case, the main effects of these predictors were in general as expected: increasing *tempo* made gestures temporally shorter, faster and smaller; gestures within sentence final *position* were of longer durations, smaller spatially and slower than their mid-sentence counterparts. *Place in the series* had no effect relevant to this study.

In some cases (C_2 duration, V_1 duration; duration of bilabial closing gesture, peak velocity of bilabial closing gesture; duration of vowel transition, peak velocity of vowel transition), the variable was transformed to a logarithmic scale to better approximate the homogeneity of variances required for ANOVA.

After first presenting the general results using ANOVA, a Bayesian analysis of the central hypothesis (9) regarding dependency of V2LAG on gemination and context will be provided.

Throughout, the results are compared with model predictions 1—9 listed at the end of Section 2. In Section 3.3.6 we provide a summary of the match (or lack thereof) between the predictions and behavior of our subjects.

3.3.1 Acoustic measurements (predictions 1,2)

Our acoustic measurements show a high degree of consistency with previous studies on Finnish summarized in the introduction.

Gemination had a significant effect on the ***acoustic duration of the bilabial constriction*** for all speakers, both in sentences and repetitions. On average, the geminates were approximately twice as long as the singletons. Vocalic and consonantal context did not have the same main effects for all speakers. There was a significant effect of *C1-context* for speaker S2 in the sentences ($p < 0.001$) with constriction longer for the preceding consonant /t/, and S1 in the repetitions in the same direction ($p < 0.05$). For speaker S3 there was a significant influence of *VV-context* for both sentences and repetitions ($p < 0.05$); in both cases, the constriction was longer for /a-i/ than /i-a/. No other main effects were significant. Although these tendencies are

speaker dependent, all significant effects and their directions are compatible with our modeling data (see prediction (1) in section 2.5).

As expected for Finnish, *gemination* had a significant lengthening effect on the ***duration of the preceding vowel***, for all speakers in both conditions ($p < 0.001$) with the exception of S3 and S4 in repetitions for which the effect was not significant. With the exception of S1 in sentences, the *VV-context* main effect was significant with /a/ longer than /i/ ($p < 0.01$ for S3 in repetitions, $p < 0.001$ for the rest). The vowel was longer when preceded by /t/ than after /p/ and this effect was generally significant.

Comparing these results with the model predictions (2), the optimal sequences differ considerably from Finnish speakers in predicting shortening of the vowel /a/ by gemination in /a-i/ context. Other observation regarding *VV-context* and *C1-context* effects match speakers' behavior.

3.3.2 Articulatory measurements: bilabial closing gesture (predictions 3—5)
The ***duration of the bilabial closing gesture*** is defined here as the duration of the interval between the onset of lip closing movement and the instant of maximal constriction. As predicted by the model (3), *gemination* had a significant effect on this measure, lengthening the gesture's duration in both sentences and repetitions ($p < 0.001$). The effect of *C1-context* on the duration of lip closing was also (with one exception, S2 in repetitions) significant, with shorter gestures in /p/-sequences ($p < 0.05$ for S4 in sentences, $p < 0.001$ for the rest). These two effects were predicted by our model (3).

Contrary to the model's predictions (3), however, our analysis found no consistent influence of *VV-context* on the duration of the closing movement. For two speakers with a significant vowel effect in the repetition task (S3 and S4) the duration was – at variance with the model – greater for /i-a/ than for /a-i/.

The ***spatial displacement*** of the lips during the closing gesture is measured as the difference between the maximal lip aperture at the onset of the gesture and its value at the maximal constriction. The spatial extent of the gesture was significantly influenced by *gemination* for all speakers in both conditions and was greater for geminates than for singletons ($p < 0.001$). *VV-context* also had a significant effect: with one exception (S1 in sentences) the extent of movement was greater for /a-i/ than for /i-a/ context ($p < 0.001$). The effect of *C1-context* is mostly significant for three out of four speakers (with an exception of S2 for sentences) but not significant for speaker S4 in either condition. In all significant cases, the extent of movement was greater when the preceding consonant was /t/.

The model predictions (4) generally match the main effects and the directions of gemination and context influence. Simulations, however, somewhat overestimate the influence of *C1-context* at the expense of *gemination* and *VV-context*.

We also analyzed the effects on the ***peak velocity*** of the closing movement. *VV-context* had a highly significant effect on the peak velocity for all speakers in both conditions ($p < 0.01$ for S1 in sentences, $p < 0.001$ for the rest). The peak closing velocity was higher for /a-i/ context. This is not surprising in the light of the observations reported above that in this context the lip closing movement spanned a greater distance within an equal or shorter duration than for the /i-a/ context.

Results are less clear for the effects of *C1-context* and *gemination* as the movements were both greater and longer for the /t/-context and for geminates (see above). The effect of *C1-context* is nevertheless significant in most cases (sentences: $p < 0.001$ for S1, $p < 0.05$ for S2, S3, S4; repetitions: $p < 0.001$ for S1 and S3), the velocity was greater for the /t/-context for all speakers except S2 and S4 in sentences, who exhibited significantly higher peak velocity in words

starting with /p/. The influence of *gemination* showed even greater speaker dependence. Except in one case (repetitions for S3) the effect was significant. *Gemination*, however, seems to act in different directions for different speakers: for S4, the peak velocity was greater for singletons than for geminates in both conditions, while for S2 it was always greater for geminates. For S1 and S3, it was greater for geminates in sentences, but greater for singletons in repetitions (not significantly so for S3).

Comparing this behavior with model predictions (5) reveals considerable differences. While our simulations predicted a strong and consistent effect of *C1-context*, the effect is less dominant in the data, although with one exception (S2, only significant in sentences) the effect was as expected: greater velocity for /t/- than for /p/-sequences. The *VV-context* effect is correctly predicted only within the /t/ consonantal context. Moreover, the data do not generally bear out the prediction of greater velocity for singletons than geminates.

Our simulations predicted a strong interaction between *C1-context* and *VV-context*, the former actually reversing the influence of the latter in different contexts. Indeed, the data analysis did find a mostly significant interaction between *C1-context* and *VV-context* is (sentences: $p < 0.001$ for S1, $p < 0.05$ for S2, S3, n.s. for S4; repetitions: $p < 0.001$ for S1, S3, $p < 0.01$ for S4, n.s. for S2). This interaction is, however, not strong enough to reverse the influence of *VV-context* in words starting with /p/. This observation further highlights the already mentioned exaggeration of the *C1-context* effect on the predicted kinematic characteristics of C_2. This is in fact solely due to the influence of the word initial /p/; our simulated /t/-sequences, on the other hand, agree with the analysis results in most qualitative aspects.

3.3.3 Articulatory measurements: transition between the flanking vowels (predictions 6—8)
Due to the difficulties of assessing the articulatory offset of the transition to the second vowel of our sequences in sentences (in particular for sentence final position), we performed the analysis only for the repetition tokens.

The ***duration of the transition*** between articulatory targets of the vowels preceding and following the bilabial showed a consistent dependence on *C1-context*. *C1-context* influence was significant for all speakers ($p < 0.001$), the transition taking longer in the /p/ context than in the /t/ context. *VV-context* was also significant ($p < 0.001$ for S1, S2, S4, $p < 0.01$ for S3), but while for speakers S1, S3 and S4 the duration was greater in the /a-i/ context as predicted by the model (prediction 6), for speaker S2 the opposite movement from /i/ to /a/ took longer.

The effect of *gemination* was also significant for all speakers ($p < 0.001$ for S1, S2, S3, $p < 0.05$ for S4). In accordance with the model prediction the vowel transition was longer in the geminate case for three speakers (S1, S2, S3), but shorter for the fourth speaker (S4).

While the model fails to account for the transition duration dependency on *C1-context*, the simulations (6) agree with the effects of *gemination* and *VV-context* prevalent among speakers.

Although the model does not account for any effects on ***displacement of tongue body*** during the inter-vocalic transition (7), we list here effects found in the data.

The distance between the extreme positions of the flanking vowels during the transition is, as expected, strongly dependent on the articulatory properties of the vowels themselves, captured here by the *VV-context* effect. The distance is greater for the /i/ to /a/ transition ($p < 0.001$). Perhaps more surprisingly, it also depends significantly on *C1-context* ($p < 0.001$ for S1, S4, $p < 0.01$ for S3), the transition being spatially greater in the /p/-context than in the /t/-context. *Gemination* also had a mostly significant influence ($p < 0.001$ for S1, S2, S4). For all speakers the distance was greater for singletons than for geminates.

The ***peak velocity of the transition*** movement seems to be the vowel transition measure most robustly influenced by *gemination*. This is not surprising, as the transition was shorter temporally (with the exception of S3) and covered a greater distance for singletons than for geminates. Indeed, the effect in the expected direction (transition faster for singletons than for geminates as predicted by the model, prediction 8) was significant for all speakers ($p < 0.001$). For speakers S1, S3 and S4 there was a significant effect of *VV-context* ($p < 0.001$) with /i-a/ transition the faster one. Again, this is to be expected, as this transition had on average a significantly greater displacement and shorter duration (speaker S2 had a significantly longer duration for /i-a/ than for /a-i/).

The durational and temporal measurements make no clear prediction for the effect of *C1-context* as the transition was mostly longer both spatially and temporally for the /p/-context. The analysis showed that the peak velocity was also significantly greater in this context for two speakers, S1 and S4 ($p < 0.001$). There were no robust interaction patterns.

Our simulations account for the peak velocity dependencies rather well (8). While failing to account for the weak influence of *C1-context*, they correctly predict the influences of *vocalic context* and *gemination*.

3.3.4 Articulatory measurements: coordination between consonant and vowel transition (prediction 9)

The measure we use to assess the coordination between the bilabial closing gesture and the vowel transition is V2LAG, defined as the time from the onset of movement towards bilabial closure up to the onset of transition between the flanking vowels. The more positive this measure is, the greater the lag of the vowel transition onset relative to the onset of the closure movement (or, equivalently, the earlier the lead of the latter relative to the former).

Our analysis showed a significant and robust influence of all three predictors of interest on the inter-gestural coordination: *gemination* ($p < 0.001$), *C1-context* ($p < 0.001$) and *VV-context* ($p < 0.05$ for S1, S3 in repetitions, $p < 0.001$ for the rest). The main effect were all in expected directions: V2LAG is greater (more positive) for geminates than for singletons, greater for the /t/ context than for the /p/ context, and greater for the /a-i/ transition than for the /i-a/ transition.

There was also an almost universally significant interaction between *C1-context* and *VV-context* ($p < 0.001$, except $p < 0.01$ for S2 in sentences and n.s. for S2 in repetitions). This interaction is due to the fact that while V2LAG is much greater for /a-i/ compared to /i-a/ in the /t/ context, the difference is not nearly so great – or is even opposite in sign – in the /p/ context (see Figure 11 in the following section).

All three strong main effects correspond to the effects of gemination and articulatory context predicted by the model (see Table 2). Although the context effects acted uniformly, there was one exception to the assumed direction of the gemination effect: in the model there was no difference in V2LAG for the pair /pipa/-/pippa/. This suggests a strong interaction between articulatory context and gemination for this measure. Also our data analysis found speaker dependent instances of such interactions, no general pattern emerged. We will investigate this issue more closely in the following section using Bayesian analysis and visualization.

3.3.5 Bayesian analysis of consonant-vowel coordination

As an additional check on the ANOVA results, we also assessed the various effects in our data with Bayesian inference using a hierarchical (or ANOVA-type) model (cf. e.g. Gelman & Hill, 2007). In this model effects corresponding to rows of a traditional ANOVA table are modeled as sets of coefficients which are normally distributed with mean and variance parameters having

non-informative prior distributions[4]. All empirical measures were included in the model simultaneously, with any logical (i.e. non-empirical) restrictions holding between them (e.g. the fact that velocity of lip movement at movement onset and movement offset is zero and less than or equal to maximum velocity everywhere in between) imposed on the prior distributions.

There are numerous advantages to using Bayesian inference: for instance there is no particular requirement of homogeneous variances, unbalanced tables present no difficulties, and above all it is relatively easy to incorporate all the relevant data into a single analysis so that there is no danger of multiple tests (Gelman *et al.*, 2012). Any inferences to be made are based on the joint posterior distribution of all parameters (i.e., unknowns, including missing data such as several independent variables which could only be measured for the repetition data). We simply calculate the posterior probability that the state of affairs in question holds true.
Alternatively, to obtain a number that more closely resembles the traditional significance of null hypothesis tests, we can calculate the posterior probability that the condition is false.
For dichotomous variables (such as singleton vs. geminate) we will use the posterior probability that the effect is opposite in direction to the median value for the coefficient in question as a measure of "significance." Thus a smaller value indicates a more significant result as usual, but unlike traditional null hypothesis testing, one minus this "significance" is also meaningful as the probability that the effect does indeed go in the specified direction.

In general the Bayesian results (shown in Table 2 for main effects) were in agreement with the ANOVA results. Here we present in more detail the results dealing with V2LAG.

Note to Publisher: Insert Figure 11 about here

To aid assessment of the overall situation features of the posterior distribution itself can be shown visually in many ways. Here we use a two dimensional representation (Figure 11) restricted to two variables at a time and show the two marginal median values for a particular group as a point, with each point accompanied by a cross-hair indicating the marginal 95 % credible intervals (95 % CI, including 95 % of the marginal posterior distribution). Many robust effects of the data can be seen fairly directly in Figure 11 (showing lip aperture at the onset of the lip closing gesture as well as the time of onset relative to V2 onset), by comparing the points and noting whether their 95 % CI overlap and to what extent. To facilitate comparison of features of the data with the simulated optimal values obtained for the model, Figure 12 shows the model results in a parallel fashion.

Note to Publisher: Insert Figure 12 about here

For instance it is evident in Figure 11 that geminates generally start earlier relative to V_2 onset (i.e. V2LAG is longer). This is also the result obtained for the model simulations (cf. prediction (9) and Table 2), as is clearly visible in Figure 12. In fact, the overall V2LAG difference for geminates compared to singletons has a posterior median value of +31.3 ms ($p < 0.0001$). The same relation holds true for all speakers in the /a-i/ words (for pappi-papi +31.1, +74.7, +33.8 and +54.8 ms, $p < 0.0001$; for tappi-tapi +27.1, +70.2, +30.4 and +50.8 ms, $p < 0.0001$, although is it larger for speakers S2 and S4, and smaller for S1 and S3 (for whom it is actually smaller than the effect of /t/ vs. /p/).

The difference is reduced considerably in the /i-a/ words for all speakers, but is still very robust for speakers S2 and S4 (pippa-pipa +36.7 and +17.0 ms, $p < 0.0001$; tippa-tipa +43.8 and +26.5 ms, $p < 0.0001$). For speakers S1 and S3 V2LAG differences in the /i-a/ words are much smaller. For tippa-tipa the difference is most likely positive (median value), but small and unreliable (S1

[4] Such a model can be seem as a principled compromise between complete pooling of data sets (ignoring some variable, such as speaker) and no pooling at all. This is sometimes referred to as "partial pooling," with the degree of pooling determined by the data.

+3.4 ms, $p = 0.1866$; S3 +7.5ms, $p = 0.0376$). For pippa-pipa the median difference is actually negative, though very small and unreliable (S1 −5.8 ms, $p = 0.0372$; S3 −3.1 ms, $p = 0.1942$). In the model results the one exception to the rule of longer V2LAG for geminates was the pair pippa-pipa (open circles in Figures 10 and 11). In addition the tippa-tipa pair had a relatively small difference. The model thus appears very much in agreement with our speakers in terms of the effect of geminates on V2LAG, and especially so for speakers S1 and S3.

The model also predicts longer V2LAG for the /t/ stimuli compared to the /p/ stimuli (also clearly visible in Figure 12). This relation is also clearly brought out by Figure 11 for the data: all t-p pairs (squares vs. circles) for all speakers show this relation (with one exception, tipa-pipa for S4), and the difference is very robust in most cases. Interestingly there appears again to be a basic difference in the pattern for S1 and S3 as opposed to S2 and S4. For S1 and S3 the C1context effect (/t/ vs. /p/) is larger than the gemination effect, whereas for S2 and S4 the reverse is the case. Here as well the model results are very much in agreement with our speakers, and especially so for S1 and S3.

Lastly the model predicts longer V2LAG for the /a-i/ words compared to the /i-a/ words in all cases. Here again the data show the predicted relation for all speakers (very robust in most cases), with one notable exception: unlike the model, speakers S1 and S3 reversed the usual direction for the pair papi-pipa, and the difference is quite robust (−27.7 and −21.9ms, $p < 0.0001$). Actually all speakers exhibited a much smaller difference in V2LAG for the papi-pipa pair, but the difference remained positive for S2 and S4, and the same holds true for the model results.

3.3.6 Comparison between prediction and empirical data

Table 2 summarizes the main effects and its directions found by data evaluation, using both ANOVA and Bayesian analysis, and the level of agreement between the empirical findings and the predictions of the model. For model predictions, symbols < and > mean strong main effect, ≤ and ≥ mark weaker and ≈ no predicted effect. For ANOVA analysis, < and > show a robust effect across (almost) all speakers/conditions, while ≤ and ≥ depict an effect that is significant for at least some cases. The results of Bayesian analysis depict main effect for an "average" speaker (partial pooling) and the symbols can be interpreted in term of significance of the effect: $p < 0.01$ (< and >), $p > 0.01$ but $p < 0.05$ (≤ and ≥), $p > 0.05$ (≈).

Checkmark in the last column (✔) signifies a reasonably close agreement and cross (✗) marks considerable disagreement (effects in opposite directions); dash (–) is used in the cases where the model hasn't predicted an effect or the prediction is only partially supported by data.

Table 2. *A summary of model predictions (column 4), the results of data analysis, both ANOVA and Bayesian analysis (columns 5 and 6), and an evaluation of the match between the predictions and empirical data. See text for details. Notes:* [*] *Speaker S4 in fact shows a predicted (>) effect;* [**] *Opposite effect (>) for S2 and S4 in sentences.*

			model	ANOVA	Bayes.	match
Acoustic dur.	C	/p/ – /pp/	<	<	<	✔
		/ai/ – /ia/	>	≥	≥	✔
		/p/ – /t/	≤	≤	<	✔
	V_1	/p/ – /pp/	> (a-i)	<	<	✗
		/ai/ – /ia/	≥	>	>	✔
		/p/ – /t/	<	<	<	✔
Bilabial	dur.	/p/ – /pp/	<	<	<	✔
		/ai/ – /ia/	>	≤	≤	✗
		/p/ – /t/	<	<	<	✔
		/p/ – /pp/	≤	<	<	✔

	displ.	/ai/ – /ia/	≥	≥	>	✔
		/p/ – /t/	<	≤	<	✔
	p.v.	/p/ – /pp/	>	< *	≈	✗
		/ai/ – /ia/	>(t) <(p)	>	>	–
		/p/ – /t/	<	≤ **	<	✔
Vowel transition	dur.	/p/ – /pp/	<	≤	<	✔
		/ai/ – /ia/	>	≥	>	✔
		/p/ – /t/	≈	>	>	–
	displ.	/p/ – /pp/	≈	>	>	–
		/ai/ – /ia/	≈	<	<	–
		/p/ – /t/	≈	>	>	–
	p.v.	/p/ – /pp/	>	>	>	✔
		/ai/ – /ia/	<	<	<	✔
		/p/ – /t/	≈	≥	>	–
V2LAG		/p/ – /pp/	≤	<	<	✔
		/ai/ – /ia/	>	>	>	✔
		/p/ – /t/	<	<	<	✔

4.0 Discussion

Overall, the optimal CV(C)CV sequences generated by our model reached a high degree of agreement with acoustic and articulatory data obtained from Finnish speakers. This agreement was mostly qualitative, it was not our aim to find a set of model parameters (defining the details of vocal tract anatomy, for example) to attempt quantitative agreement. The main focus was on the effects of consonant quantity and articulatory context on inter-gestural coordination. To evaluate the model performance more thoroughly, we also compared the influence of these variables on kinematic properties of the bilabial gesture undergoing gemination and the coproduced articulatory transition between the flanking vowels.

It is important to note here, that the optimal gestural scores used for predicting context dependency and effects of gemination were obtained using the same optimization procedure with the same set of parameters except, of course, the local premium used to elicit gemination. The segmental differences between sequences come purely from different "naïve" gestural scores fed to the optimization procedure (akin to the score shown in Figure 2). These inputs determine what segments – realized articulatory gestures – are to be present in the optimal sequence and in what order *and nothing more*. It is a basic property of the optimization process, that it arrives at the same solution minimizing the given cost function regardless of the starting point.

In this sense, the characteristics of individual gestures in the resulting optimal sequences, inter-gestural relations and temporal characteristics of constrictions (with the exception of geminates being longer than singletons elicited by the local premium) can be seen as *emergent* from the spatial and "physical" characteristics of the model vocal tract and the trade-offs between production, perception and high-level temporal constraints as implemented in the cost function underlying the model behavior. In short, the model has not received any explicit phonological knowledge (apart from the fact that geminates are longer that singletons) or any phonetic information derived from data (e.g., that /a/ is generally longer than /i/).

Obviously the question arises how various settings of the model influence the qualitative and quantitative aspects of phenomena discussed here. Can the optimization platform presented in this work predict patterns that directly contradict the predictions presented here? In short, can the model predict anything given the right parameters? Although there is not space here for a thorough analysis of this important issue, our ongoing investigation suggests that the short

answer is: No[5]. The parameters can obviously influence quantitative aspects of the optimal sequences. Given appropriate settings, the articulators can move infinitely slowly or over vast distances (or even extremely fast over vast distances). However, provided that the main structural characteristics – basic anatomical links, realistic relationship between masses of the articulators, nature of tract variables and gestural targets – of the model remain intact, the parameter tuning cannot reverse the reported relationship between, for example, the closure duration or consonantal context and inter-gestural timing.

At the same time, the model can be individualized. As we saw, the modeling predictions match the behavior of some speakers better than others. This might be due to physiological differences between our subjects – incidentally the model accounted better for the interaction patterns exhibited by two female speakers (S1 and S3) compared to the males (S2 and S4). Changing the settings of the model that correspond to the individual physiological characteristics of the speaker, such as dimensions of the vocal tract and the masses of articulators, might lead to a slight shift in its behavior towards the interaction patterns exhibited by the male speakers. This would help further elucidate the nature of articulatory coordination as embodied efficient action.

Some predictions and corresponding behavior of human speakers, in particular several of the effects of gemination, are quite straightforward consequences of the fact that geminate consonants are longer than singletons. Among these are the observations that geminate consonantal gestures are longer and spatially larger than singleton ones.

Many aspects of the model's predictions can be accounted for as "common-sense" rational solutions of the task at hand in the given circumstances; that is, after all, the underlying nature of optimization. An example could be the effect of consonantal context on the kinematics of the C_2 gesture. When the bilabial C_2 is preceded by an alveolar gesture that has only a small influence on the lip movement, the lips are generally further apart when starting their movement towards each other than when the preceding consonant is another bilabial[6]. The spatial extent of the gesture can thus be expected to be smaller in /p/ than in /t/ context. Similarly, when the preceding vowel is /i/ whose rising tongue movement pushes the jaw and the lower lip upwards, the lip closure is assisted by the vowel's articulation and can be temporally shorter than in /a/ context (Šimko *et al*., 2011).

In both consonantal contexts explored here, the upwards push of the jaw for consonants contributes to faster movement of the tongue body towards its gestural target; consequently, the optimal duration of /i/ in V_1 position is shorter than that of /a/. Finally, lower tongue position during coproduction of V_1 with /t/ than with /p/ (see Figures 8 and 9) results in the correctly predicted longer V_1 in /t/ than in /p/ context. As these patterns reflect physiology of the tongue-jaw-lips system, they require the type of embodied modeling used here capable of capturing complex synergies among speech articulators.

Another group of phenomena that our model predicted successfully could be a fortuitous consequence of the model's architecture. Vowel-to-vowel transition is longer and slower in the geminate context in both predictions and recordings. As the spatial extent of the transition in our simulations is (incorrectly) invariant, the differences may be attributable to lower overall stiffness of the optimal sequences with geminates compared to those with singletons, see Figures 5 and 6. Similarly, the correctly predicted lower peak velocities of the transition for /a-i/ compared to /i-a/ context presumably arise from the stiffness differences. As the stiffness of

[5] In a study currently being prepared for publication we investigate whether the model can be re-parameterized to exhibit the observed differences between speakers. Testing model's behavior for individual parameters varied by as much as 30 % of the values used in the present work shows that the patterns presented here are robust for parameter values within realistic ranges.

[6] It is also much more difficult (for adults) to rapidly repeat a series of syllables all starting with /p/ than to repeat syllables which alternate between /p/ and /t/. For Finnish see Lehtonen (1971).

individual gestures in our model is always proportional to the overall stiffness parameter being optimized, the model cannot account for possible *relative* changes of stiffness of individual gestures in various contexts. Although the predictions are correct, we cannot be sure that they are correct for the right reasons.

We assume that the reported oversensitivity of the optimal sequences to consonantal context is also a result of limitations of the model's design. As its gestural repertoire does not include an active lip opening gesture the lip opening is driven purely by speech-ready dynamics and is comparatively sluggish. Consequently, the lip opening following bilabial consonant /p/ is considerably smaller than after alveolar /t/. Although this seems to be also true for our speakers (closures larger in the /t/ context than in the /p/ context, significantly so for three speakers), the extent of the effect is much smaller. Adding an active lip opening gesture and a lip aperture measure as a contributor to vowel quality in our model might correctly mitigate the exaggeration of the consonantal context effect (see Beňuš & Šimko (this issue) on incorporating an active opening gesture in a simplified version of the ETD model).

Data analysis of articulatory kinematics of the bilabial and the coproduced vocalic gestures revealed that each of the model's predictions corresponded quite well with two out of three kinematic characteristics under investigation. The characteristics we evaluated were duration, displacement and peak velocity of the given gesture. For the bilabial closing gesture, the model accounted satisfactorily for the effects on duration (apart from vocalic effect) and displacement but its predictions differed quite considerably with regard to velocity. For the vowel transition, the effects on peak velocity and duration were to a large extent predicted correctly but effects on displacement were not predicted at all. A possible source of this shortcoming may be the type of dynamics –critically damped mass-spring – used in the model. The dynamics lawfully binds the three kinematic characteristics. If, as suggested by e.g. Fuchs *et al*. (2011), the critically damped second order dynamics does not provide an accurate estimate of gestural action of the human speech apparatus, the model might not be able to account for phenomena related to all three characteristics simultaneously (see also Birkholz *et al*. (2011) on the issue of the appropriate order of dynamical system for speech).

We hasten to say that the simplifications in the model architecture and dynamics mentioned here are not mere omissions. The primary reason for these simplifications of the model is technical: optimization requires tens of thousands of gestural score evaluations, and adding additional elements or more complex dynamical definitions of gestures slows the process down exponentially. Second, the ability to generate valid predictions with as simple a platform as possible confers greater explanatory power on our findings as fewer parameters allow for better identification of sources of various patterns as outlined above. It is interesting that the model produces realistic temporal coordination patterns despite these simplifications.

Interestingly, the simulations predicted no effect on V_1 duration for gemination in /i-a/ sequences, but a shortening effect in the /a-i/ context. So, in one context the model behaves almost like our Finnish speakers and in another like an Italian or a Swede. We offer no explanation for this phenomenon. In any case, we imposed no particular known pattern of durational sensitivity of the vowel to quantity of the following consonants. The fact that model arrived at both of these patterns – albeit with no V_1 lengthening and in two separate vocalic contexts – may suggest two stable solutions for inter-gestural timing instantiated in two broad groups of languages. Provided that the solutions do not differ considerably in their respective impact on efficiency requirements as utilized here, speakers of different languages may opt for one or the other depending on other factors, for example the rhythmical properties of their language. The reason why no languages seem to use the context dependent coordination as predicted by our model might be related to another influence considered by Lindblom et al. (2011) – the learning cost related to generalizability of acquired patterns.

Several of the context dependent patterns mentioned above (e.g. spatial and temporal extent of gestures in various vocalic and consonantal contexts) can be seen as passive consequences of vocal tract physiology and the articulatory task performed. They arise naturally both in an embodied modeling paradigm such as the one presented here and in the embodied and complex human vocal tract. The relative phasing of onsets of largely overlapping vocalic and bilabial gestures – captured here by the V2LAG measure – may, however, be considered a hallmark of an active control behind the communication task at hand. The modeling results as well as the analysis of articulatory recordings strongly suggest that the details of inter-gestural coordination are influenced by requirements of production efficiency and perception efficacy.

The fact that the bilabial gesture is spatially smaller after /i/ than after /a/ may result from physiological linkages between the lips, tongue and the jaw (as instantiated in the model) or from different lip aperture targets for the two vowels. It is common sense, that in order to achieve an appropriate timing of the *effects* of the bilabial action determined indirectly by perception constraints embedded in the parsing cost component, the lips could start the closing movement relatively late in /i-a/ compared to /a-i/ (cf. Šimko *et al*., 2011). Or, for reasons similar to those mentioned above, later when the preceding consonant is /p/ rather than /t/. The model simulations do this because they are optimal. Importantly, as our analysis of articulatory recordings suggests, human subjects behave, at least statistically, in the same optimal fashion. In our opinion, this finding has strong repercussions for our understanding of articulatory control in speech and skilled cognitive action in general.

A different type of emergent phenomenon presented in this work is the bifurcation in inter-gestural coordination that emerges when a cost function parameter inducing longer closure durations reaches a certain value. For a short interval of the parameter values there exist two solutions, each a local minimum of the cost function, one with the characteristics of a singleton and the other geminate-like. Beyond this interval, on both sides, the optimal inter-gestural patterns quickly reach relatively stable constellations. Coordination patterns in these stable areas retain the singleton and geminate attributes, respectively.

The relatively stable areas with a sharp non-linear transition between them is reminiscent of the non-linearities in the articulation-to-auditory mapping claimed by Stevens' (1989) quantal theory to influence the partitioning of various phonetic continua, such as the vowel space and consonant place of articulation, into phonological categories. In the present account, the patterning is a result of similar non-linear relations between timing and embodied articulation. A smooth change in perceptually motivated requirements – prominence expressed in terms of duration – results in an abrupt change in inter-gestural coordination. Just as vowels belonging to the same plateau in Stevens' account sound different from vowels from another plateau across the non-linearity in the articulation-to-acoustic mapping, the production of consonants belonging to our two plateaux "feels" different. That is, the seemingly continuous space is experienced as discretized in an embodied fashion.

We stop short of claiming that precisely this modeling phenomenon realistically reflects the way in which the phonological quantity contrast has diachronically risen in various languages. What our modeling suggests is that that optimization based dynamical modeling, although *continuous* in its nature, can at least in principle lead to identifying *discrete* categories in the space of possible articulatory patterns. These qualitatively different patterns can subsequently serve as an affordance for encoding phonological contrasts that can be utilized for communication.

The affordance is presented in the form of distinct local minima of the cost functions depicting trade-offs between productions and perception aspects of speech. These local minima can be interpreted in terms of competing dynamical attractors that arise and disappear with continuous variation of the intentional parameter of local premium. Our approach can thus be straightforwardly recast in terms of non-linear dynamics that has been successfully used to model coexistence of the continuous and the discrete in terms of attractors determining

qualitative behavior of complex continuous systems (Haken et al., 1985; Gafos, 2006). In addition, the optimization paradigm provides a link between the attractor landscape – determined by the shape of the composite cost function – and the properties of the underlying embodied production and perception apparatuses.

5.0 Conclusions

Analysis of articulatory recordings of two-syllable $C_1V_1C_2V_2$ sequences spoken by Finnish speakers revealed a substantial degree of dependency of kinematic characteristics of articulator movement and inter-gestural coordination on (1) consonantal quantity (singleton vs. geminate) of the bilabial consonant C_2, (3) the articulatory nature of the transition between the vowels and (3) the extent to which C_1 interferes with articulation of the following consonant. These influences, moreover, show rather complex interaction patterns. Articulatory kinematics reflects spatial and temporal attributes of sequenced and coproduced gestures and synergistic effects among the coordinated articulators.

The kinematic properties and coordination patterns revealed by our analysis show a high degree of agreement with predictions of the optimization-based embodied task dynamical model. In the model, gemination was elicited through local adjustments of a premium placed on perceptual properties of a consonant in terms of its relative prominence. The model generates its predictions as realizations of a given task that are optimal with respect to trade-offs between competing requirements of production efficiency and perceptual clarity.

In the model, phasing relations between gestures are a part of an efficient solution reflecting physiological constraints of the embodied speech apparatus as well as the communication task. The match between the model's predictions and the behavior of speakers provides strong support for the hypothesis that speech articulation is subject to the same efficiency requirements as those guiding many other types of skilled target-oriented action. The agreement between predictions and data was particularly high in the case of details and context-dependency of inter-gestural timing that is usually considered actively controlled and learned as a phonologically specified coordination mechanism.

The model also predicts discretely distinct coarticulation patterns distinguishing singleton and geminate production, highlighting the ability of a continuous dynamical optimization-based approach to account for emergent qualitative contrasts. The predicted differences in inter-gestural phasing underlying the emergence of contrast were identified by analysis of empirical data. To confirm or refute the manner in which the qualitative differences emerges in our model will, however, require further experimental research primarily focusing on elicitation of the consonantal quantity contrast by continuous means.

Acknowledgements

This work was partly supported by an Alexander von Humboldt Fellowship grant to the first author. We would like to thank Mona Lehtinen for her help with articulatory recordings and data processing. We are also indebted to the four speakers who participated in this work. Finally, we are grateful to two anonymous reviewers for many constructive comments and suggestions that helped improve the paper considerably.

Appendix

The following tables complement our report on the results of statistical analysis (ANOVA) of the empirical data presented in Sections 3.3.1—3.3.4. *F*-values for main effects of *gemination*, *VV-context* and *C1-context* as well as for interactions between these three variables are listed. In the interest of brevity, we omit *F*-values for other variables (*position*, *tempo*, *place in series*) that were also incorporated in ANOVA analyses.

Significance of the effect is marked by the asterisks in the usual way: $p < 0.05$:*, $p < 0.01$:**, $p < 0.001$:***. Significant effects are highlighted in bold. Table captions refer the relevant model prediction.

Table A1. *Acoustic duration of bilabial constriction (prediction 1).*

	S1		S2		S3		S4	
	sent's *F*(1,94)	rep's *F*(1,156)	sent's *F*(1,135)	rep's *F*(1,72)	sent's *F*(1,145)	rep's *F*(1,7)	sent's *F*(1,156)	rep's *F*(1,80)
gem	**1061*****	**1902*****	**1835*****	**3492*****	**1413*****	**2118*****	**876*****	**609*****
V-cont	*2.24*	*0.23*	*0.08*	*0.16*	**5.89***	**8.25***	*2.02*	*3.52*
C_1-cont	*0.82*	**6.75***	**11.4*****	*2.64*	*0.91*	*4.22*	*0.80*	*0.48*
gem:V	*0.01*	*1.11*	*0.74*	**19.3*****	*0.04*	*0.15*	*2.04*	**6.16***
gem:C_1	*0.09*	*0.39*	*2.06*	**6.19***	*1.02*	**12.1***	*2.13*	*1.01*
V:C_1	*1.23*	*0.40*	*0.20*	**6.57***	*2.64*	*0.96*	*0.06*	*2.32*
gem:V:C_1	*0.44*	*0.01*	*0.05*	**10.20****	*0.18*	**6.37***	*1.93*	*1.87*

Table A2. *Acoustic duration of the vowel preceding the bilabial (prediction 2).*

	S1		S2		S3		S4	
	sent's *F*(1,94)	rep's *F*(1,156)	sent's *F*(1,135)	rep's *F*(1,72)	sent's *F*(1,145)	rep's *F*(1,7)	sent's *F*(1,156)	rep's *F*(1,80)
gem	**77.1*****	**41.5*****	**27.4*****	**42.9*****	**35.9*****	3.44	**171*****	1.32
V-cont	*1.63*	**68.2*****	**75.0*****	**230*****	**11.6*****	**10.8***	**11.5*****	**29.6*****
C_1-cont	*2.62*	**11.4*****	**8.94****	**63.5*****	**6.93****	**12.0***	*1.03*	**22.6*****
gem:V	*0.64*	**6.79***	*0.09*	*0.49*	*0.05*	**28.9****	*1.71*	*3.61*
gem:C_1	*0.05*	*0.06*	*0.03*	**4.88***	*0.25*	*4.81*	*0.01*	*1.30*
V:C_1	*0.03*	*0.06*	*0.30*	**4.82***	**4.83***	*2.82*	*0.10*	**7.02***
gem:V:C_1	*0.35*	*0.99*	*0.37*	*0.65*	*0.14*	*0.45*	*1.48*	*0.44*

Table A3. *Duration of the bilabial closing gesture (prediction 3).*

	S1		S2		S3		S4	
	sent's *F*(1,94)	rep's *F*(1,156)	sent's *F*(1,135)	rep's *F*(1,72)	sent's *F*(1,145)	rep's *F*(1,7)	sent's *F*(1,156)	rep's *F*(1,80)
gem	**514*****	**551*****	**169*****	**66.3*****	**396*****	**737*****	**642*****	**1161*****
V-cont	*0.59*	*0.81*	*3.86*	*0.90*	*2.69*	**15.9****	*0.84*	**11.9*****
C_1-cont	**82.1*****	**113*****	**15.2*****	*0.04*	**16.1*****	**301*****	**4.10***	**88.0*****
gem:V	**6.69***	*0.83*	**5.33***	*1.86*	*0.49*	**19.1****	*0.03*	*0.55*
gem:C_1	*0.04*	*1.80*	**4.98***	**5.95***	*0.21*	**31.1*****	*1.11*	**10.7****
V:C_1	*0.06*	**8.90****	*0.07*	**6.45***	*0.10*	*4.27*	*1.58*	**4.77***
gem:V:C_1	*0.09*	*0.00*	*0.08*	**16.9*****	*0.10*	**8.68***	*0.27*	*0.25*

Table A4. *Spatial displacement of the bilabial closing gesture (prediction 4).*

	S1		S2		S3		S4	
	sent's *F*(1,94)	rep's *F*(1,156)	sent's *F*(1,135)	rep's *F*(1,72)	sent's *F*(1,145)	rep's *F*(1,7)	sent's *F*(1,156)	rep's *F*(1,80)
gem	**85.2*****	**24.2*****	**28.4*****	**16.0*****	**108*****	**161*****	**68.6*****	**59.8*****
V-cont	*3.55*	**115*****	**102*****	**495*****	**102*****	**135*****	**83.1*****	**16.6*****
C_1-cont	**55.6*****	**321*****	*1.18*	**9.42****	**7.89****	**538*****	*0.17*	*3.38*
gem:V	*1.19*	**5.04***	*0.10*	*0.72*	*0.00*	**22.2****	*0.24*	*0.50*
gem:C_1	*1.82*	*2.13*	*1.43*	**8.10****	*2.70*	*4.44*	*0.73*	*0.58*
V:C_1	**4.28***	**20.4*****	2.48	0.66	**4.23***	**10.8***	3.88	**4.16***
gem:V:C_1	*0.10*	*0.87*	*0.15*	*1.11*	*0.26*	**25.4****	**4.29***	*0.61*

Table A5. *Peak velocity of the bilabial closing gesture (prediction 5).*

	S1	S2	S3	S4

	sent's *F*(1,94)	rep's *F*(1,156)	sent's *F*(1,135)	rep's *F*(1,72)	sent's *F*(1,145)	rep's *F*(1,7)	sent's *F*(1,156)	rep's *F*(1,80)
gem	**6.99****	**41.4*****	**6.41***	**17.8*****	**13.4*****	*0.65*	**7.44****	**12.7*****
V-cont	**6.62***	**113*****	**137*****	**224*****	**137*****	**403*****	**76.9*****	**22.7*****
C_1-cont	**38.4*****	**138*****	**5.33***	*1.68*	**6.10***	**439*****	**4.17***	*0.51*
gem:V	**5.55***	**5.53***	*0.02*	*0.19*	*1.97*	**14.7****	*0.00*	*0.05*
gem:C_1	*2.23*	*1.02*	*0.12*	**8.41****	**5.91***	*1.49*	**4.20***	*0.03*
V:C_1	**14.8*****	**26.9*****	**4.47***	*0.43*	**6.17***	**65.9*****	*0.76*	**7.24****
gem:V:C_1	*0.02*	*0.88*	*0.49*	*1.55*	*0.86*	**19.1****	**6.67***	**4.46***

Table A6. *Duration of the intevocalic transition (prediciton 6).*

	S1 rep's *F*(1,141)	S2 rep's *F*(1,62)	S3 rep's *F*(1,6)	S4 rep's *F*(1,72)
gem	**369*****	**43.4*****	**113*****	**4.35***
V-cont	**96.3*****	**112*****	**21.7****	**38.8*****
C_1-cont	**18.0*****	**17.9*****	**40.3*****	**19.0*****
gem:V	*0.12*	**54.9*****	*0.13*	**22.9*****
gem:C_1	*1.84*	**10.3****	*0.27*	**28.1*****
V:C_1	*0.80*	*0.15*	*2.44*	*2.44*
gem:V:C_1	*0.01*	**5.69***	*0.48*	**14.9*****

Table A7. *Tongue body displacement during the intevocalic transition (predicition 7).*

	S1 rep's *F*(1,141)	S2 rep's *F*(1,62)	S3 rep's *F*(1,6)	S4 rep's *F*(1,72)
gem	**75.5*****	**12.8*****	*2.35*	**62.3*****
V-cont	**14.4*****	**47.7*****	**45.1*****	**166*****
C_1-cont	**41.0*****	*3.25*	**22.8****	**90.7*****
gem:V	*0.77*	*0.02*	*0.83*	**9.42****
gem:C_1	**5.23***	**24.1*****	**15.6****	*2.99*
V:C_1	**13.2*****	*1.65*	**39.6*****	**49.9*****
gem:V:C_1	*0.92*	*3.76*	**8.02***	**5.04***

Table A8. *Peak velocity of the intevocalic transition (predicition 8).*

	S1 rep's *F*(1,141)	S2 rep's *F*(1,62)	S3 rep's *F*(1,6)	S4 rep's *F*(1,72)
gem	**1036*****	**66.0*****	**107*****	**76.9*****
V-cont	**178*****	*0.04*	**98.5*****	**481*****
C_1-cont	**32.0*****	*0.25*	*3.66*	**89.4*****
gem:V	**6.37***	*0.14*	*0.14*	*0.09*
gem:C_1	*0.00*	**21.4*****	*2.87*	*1.04*
V:C_1	*0.02*	*0.28*	*0.15*	**41.8*****
gem:V:C_1	*0.00*	**15.5*****	*0.04*	*0.89*

Table A9. *Vowel-consonant coordination: V2LAG*

	S1		S2		S3		S4	
	sent's *F*(1,94)	rep's *F*(1,156)	sent's *F*(1,135)	rep's *F*(1,72)	sent's *F*(1,145)	rep's *F*(1,7)	sent's *F*(1,156)	rep's *F*(1,80)
gem	**53.0*****	**106*****	**140*****	**670*****	**33.7*****	**41.9*****	**230*****	**193*****
V-cont	**13.0*****	**5.75***	**93.3*****	**816*****	**12.9*****	**8.41***	**271*****	**113*****
C_1-cont	**294*****	**399*****	**53.2*****	**59.5*****	**83.0*****	**104*****	**11.6*****	**55.0*****
gem:V	*0.09*	**8.55****	**68.9*****	**464*****	*0.29*	*0.23*	**60.3*****	*0.54*
gem:C_1	*0.29*	*3.50*	*0.57*	**4.47***	*0.94*	**15.0****	*3.45*	*2.16*
V:C_1	**20.3*****	**33.9*****	**8.43****	*0.12*	**31.0*****	**21.5****	**55.7*****	**129*****

gem:V:C_1	*2.49*	*0.62*	*1.60*	*0.05*	*0.02*	*1.01*	**18.9*****	**10.7****

Reference List

Kingston, J. & Beckman, M. E. (1990). Introduction. In Kingston, J. & Beckman, M. E., editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. 1-16, Cambridge: Cambridge University Press.

Beňuš, Š. & Šimko, J. (this issue). Emergence of prosodic boundary: continuous effects of temporal affordance on inter-gestural timing, *Journal of Phonetics.*

Best, C. T., Morrongiello, B. & Robson, R. (1981) Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, 29(3):191–211.

Birkholz, P., Kröger, B. J. & Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant–vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1422–1433.

Bouarourou, F., Vaxelaire, B., Ridouane, R., Hirsch, F., and Sock, R. (2008). Gemination in Tarifit Berber: X-ray and acoustic data. In Sock, R., Fuchs, S., and Laprie, Y., editors, *Proceedings of the 8th International Seminar on Speech Production*, 117–120, Strasbourg, France.

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

Elert, C. (1964). *Phonologic studies of quantity in Swedish: based on material from Stockholm speakers*, volume 27. Almqvist & Wiksells, Stockholm.

Engstrand, O. & Krull, D. (1994). Durational correlates of quantity in Swedish, Finnish and Estonian: Cross-language evidence for a theory of adaptive dispersion. *Phonetica*, 51:80–91.

Esposito, A. & Di Benedetto, M. (1999). Acoustical and perceptual study of gemination in Italian stops. *The Journal of the Acoustical Society of America*, 106:2051.

Fintoft, K. (1961). The duration of some Norwegian speech sounds. *Phonetica*, 7:19–39.

Gafos, A. I. (2006). Dynamics in grammar: Comment on Ladd and Ernestus & Baayen. In Goldstein, L., Whalen, D. H. and Best, C. T. (editors) *Laboratory Phonology 8: Varietes of Phonological Competence*, 8:51–79. Mouton Gruyter de Gruyter, Berlin, New York.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multi- level/Hierarchical Models*. Cambridge University Press.

Gelman, A., Hill, J. & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5:189–211.

Gili Fivela, B. & Zmarich, C. (2005). Italian geminates under speech rate and focalization changes: Kinematic, acoustic, and perception data. In *Proceedings of Interspeech 2005*, 2897–2900, Lisbon, Portugal.

Gili Fivela, B., Zmarich, C., Perrier, P., Savariaux, C., Tisato, G. (2007). Acoustic and kinematic correlates of phonological length contrast in Italian consonants. In *Proceedings of the XVIth International Congress of Phonetic Sciences*, Saarbrücken, Germany.

Fuchs, S., Perrier, P., & Hartinger, M. (2011). A critical evaluation of gestural stiffness estimations in speech production based on a linear second-order model. *Journal of Speech, Language and Hearing Research*, 54(4):1067–1076.

Haken, H., Kelso, J. A. S. and Bunz, H. (1985) A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51:347–356.

Homma, Y. (1981). Durational relationship between Japanese stops and vowels. *Journal of Phonetics*, 9(3):273–281.

Hoole, P. & Zierdt, A. (2010). Five-dimensional articulography. *Speech motor control: New developments in basic and applied research*, 331–349.

Kingston, J., Kawahara, S., Chambless, D., Mash, D., & Brenner-Alsop, E. (2009). Contextual effects on the perception of duration. *Journal of Phonetics*, 37(3):297–320.

Lehtonen, J. (1970). *Aspects of Quantity in Standard Finnish*. Studia philologica Jyväskyläensia, volume 6. University of Jyväskylä.

Lehtonen, J. (1971). *Diadockokinesia ja puheen temporaalinen jäsentyminen*. Jyväskylän yliopiston suomen kielen laitoksen julkaisuja 3, Jyväskylän yliopisto. (summary: Diadockokinesis and the temporal organization of speech).

Lehtonen, J. (1979). On labial co-articulation. In Hurme, P., editor, Fonetiikan päivät, Jyväskylä 1978 / *Papers from the Eight Meeting of Finnish Phoneticians*, number 18 in Jyväskylän yliopiston suomen kielen ja viestinnän laitoksen julkaisuja / Publications from the Institute of Finnish Language and Communication, University of Jyväskylä, 99–106.

Lindblom, B. (1987). Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance. In *Proceedings of the XIth International Congress of Phonetic Sciences*, volume 3, pages 9–18, Tallin, Estonia.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers.

Lindblom, B. (1999). Emergent phonology. In *Proc. 25th Annual Meeting of the Berkeley Linguistics Society*, U. California, Berkeley.

Lindblom, B., Diehl, R., Park, S., & Salvi, G. (2011). Sound systems are shaped by their users: The recombination of phonetic substance. In Clements, G. N. and Ridouane, R., editors, *Where Do Phonological Features Come From? Cognitive, physical and developmental bases of distinctive speech categories*. 67–97, John Benjamins Publishing Company.

Löfqvist, A. (1996). Control of oral closure and release in bilabial stop consonants. In P. P. McCormack & A. Russel (Eds.), *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*, 561–566, Canberra: The Australian Speech Science and Technology Association.

Löfqvist, A. (2005). Lip kinematics in long and short stop and fricative consonants. *The Journal of the Acoustical Society of America*, 117(2):858.

Löfqvist, A. (2006). Interarticulator programming: Effects of closure duration on lip and tongue coordination in Japanese. *The Journal of the Acoustical Society of America*, 120(5 Pt 1):2872.

Löfqvist, A. (2007). Tongue movement kinematics in long and short japanese consonants. *The Journal of the Acoustical Society of America*, 122(1):512.

Löfqvist, A. & Gracco, V. L. (1997). Lip and jaw kinematics in bilabial stop consonant production. Journal of Speech, Language, and Hearing Research, 40:877–893

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

Moon, S.-J. & Lindblom, B. (2003). Two experiments on oxygen consumption during speech production: vocal effort and speaking tempo. In *Proceedings XVth International Congress of Phonetic Sciences*, Barcelona, Spain.

Nam, H., Goldstein, L., Saltzman, E. and Byrd, D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *The Journal of the Acoustical Society of America*, 115:2430–2430.

O'Dell, M. (2003). *Intrinsic Timing and Quantity in Finnish*. Acta Universitatis Tamperensis 979. Tampere University Press.

O'Dell, M., Šimko, J., Nieminen, T., Vainio, M., and Lehtinen, M. (2011a). Timing of intervocalic consonant gestures in Finnish. In Werner, S. and Kinnunen, T., editors, XXVI Fonetiikan päivät 2010, pages 16–21. Itä-Suomen yliopisto.

O'Dell, M., Šimko, J., Nieminen, T., Vainio, M. & Lehtinen, M. (2011b). Relative timing of bilabial gesture in Finnish. In *Proceedings of the 17th International Congress of Phonetic Sciences*, Hongkong.

Öhman, S. E. G. (1966). Coarticulation in VCV Utterances: Spectrographic Measurements. *The Journal of the Acoustical Society of America*, 39(1):151–168.

Port, R., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America*, 81(5):1574–1585.

Pouplier, M. (2012). The gaits of speech: Re-examining the role of articulatory. In Solé, M.-J. & Recasens, D., editors, *The Initiation of Sound Change: Perception, production, and social factors*, volume 323 of Current Issues in Linguistic Theory, 147–164. John Benjamins Publishing.

Ridouane, R. (2010). Geminates at the junction of phonetics and phonology. In Fougeron, C., Kühnert, B., D'Imperio, M., and Vallée, N., editors, *Laboratory Phonology 10*, 61–90. De Gruyter Mouton.

Saltzman, E. L. & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, 1(4):333–382.

Šimko, J. (2009). *The Embodied Modelling of Gestural Sequencing in Speech*. Phd thesis, University College Dublin, Ireland.

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 the 17th International Congress of Phonetic Sciences*, 1850–1853, Hongkong.

Smith, C. (1995). Prosodic patterns in the coordination of vowel and consonant gestures. *Papers in Laboratory Phonology IV, Phonology and phonetic evidence. CUP*, 205–222.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17:3–45.

Zeroual, C., Hoole, P., and Gafos, A. I. (2008). Spatio-temporal and kinematic study of Moroccan Arabic coronal geminate plosives. In Sock, R., Fuchs, S., and Laprie, Y., editors, *Proceedings of the 8th International Seminar on Speech Production*, pages 135–138, Strasbourg, France

Figure captions

Figure 1.
Schematic picture of model anatomy depicting the model articulators and the anatomical links among them modeled as critically damped mass-springs.

Figure 2.
An example of a (non-optimal) gestural score and associated movements of model articulators computed by embodied task dynamics. Realization intervals of individual gestures are also depicted.

Figure 3.
Definition of parsing cost for individual consonants as a non-linear function of the duration of realization interval. Decreasing the slope elicits longer durations of individual gestures in the optimal gestural scores. This technique has been used to model the quantity contrast between singletons and geminates. See text for details.

Figure 4.
Optimal /api/ sequences containing "singleton" (left) and "geminate" (right) bilabial stop. The upper panes in both figures show the optimal gestural score and the lower panes the corresponding trajectories of model articulators.

Figure 5.
Optimal /ipa/ sequences containing "singleton" (left) and "geminate" (right) bilabial stop.

Figure 6.
Relationship between (A) the value of local premium placed on the bilabial stop and the synchronization measure V2LAG, and (B) V2LAG and closure duration closure CLDUR in optimal sequences /api/.

Figure 7.
Relationship between (A) the value of local premium placed on the bilabial stop and the synchronization measure V2LAG, and (B) V2LAG and closure duration closure CLDUR in optimal sequences /ipa/.

Figure 8.
Optimal gestural scores and trajectories for sequences /Cap(p)i/, C is /t/ or /p/.

Figure 9.
Optimal gestural scores and trajectories for sequences /Cip(p)a/, C is /t/ or /p/.

Figure 10.
Labeling of articulatory data. See text for description.

Figure 11.
Lip closing gesture onsets (posterior median values; cross-hairs indicate marginal 95 % credible intervals).

Figure 12.
Lip closing gesture onsets as simulated by the model.