Sequencing and optimization within an embodied task dynamic model

Juraj Simko*and Fred Cummins[1]

September 6, 2010

UCD School of Computer Science and Informatics
University College Dublin
t: +353 1 716 2902
f: +353 1 269 7262
e: *juraj.simko@uni-bielefeld.de, fred.cummins@ucd.ie*
Suggested running head: Optimization within embodied task dynamics
Keywords: Task Dynamics, Articulatory Phonology, Embodiment, Coordination, Sequencing, Optimization, Motor Control

Preprint of article to appear in *Cognitive Science*, 2010

[1] To whom correspondence should be addressed.

*Now at Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, DE

Abstract

A model of gestural sequencing in speech is proposed that aspires to producing biologically plausible fluent and efficient movement in generating an utterance. We have previously proposed a modification of the well-known task dynamic implementation of articulatory phonology such that any given articulatory movement can be associated with a quantification of effort (Simko and Cummins, 2010). To this we add a quantitative cost that decreases as speech gestures become more precise, and hence intelligible, and a third cost component that places a premium on the duration of an utterance. Together, these three cost elements allow us to algorithmically derive optimal sequences of gestures and dynamical parameters for generating articulator movement. We show that the optimized movement displays many timing characteristics that are representative of real speech movement, capturing subtle details of relative timing between gestures. Optimal movement sequences also display invariances in timing that suggest syllable-level coordination for CV sequences. We explore the behavior of the model as prosodic context is manipulated in two dimensions: clarity of articulation and speech rate. Smooth, fluid, and efficient movements result.

1. Introduction

Fluidity is the prime hallmark of biological movement, whether it be swinging through the treetops, or, as here, sequencing articulatory movements in speaking. Any child can recognize and even imitate the ratchet-like movements of a 1980's robot, but once they are done, they will stand up, walk, run, climb, pick things up, and perhaps throw them, all with a grace and lack of self-consciousness that is their biological birthright. In doing so, a sequence of behavioral goals will be reached by employing smooth movements that often overlap in time. This requires the harnessing of a massively redundant neuro-biomechanical system in the service of those goals. Because of the complexity and high-dimensionality of the embodied system employed, the goals themselves do not sufficiently constrain movement trajectories, or the sequence of forces used to generate them. Once past infanthood, the child is a skilled actor, and his movement bears the signature of his own solution to the problem of coordination. We can all write similar sequences of words, but each such performance will also be immediately attributable to this or that individual through the idiosyncrasies of handwriting. What is left unspecified by the behavioral goal is provided by the history of motor practice, gradually shaping the action space of an individual through repetition and optimization.

Not all the optimization is done during the development of an individual, however. No matter how much I practice, my skill at jumping through the treetops will never approach that of most other primates. The starting point for skill acquisition in an individual is itself the result of a long process of adaptation and change. The physical properties of the limbs, their masses, joint characteristics, their relation to the torso, these may all likewise be viewed as the result of a process of optimization carried out over an evolutionary time span, and attributable to the dynamic interplay between the organism and its environment. As I dust myself off, I may take comfort from the fact that my primate friends will never match my performance in speaking, which in my customary environment is a rather more useful skill.

We wish to advance a model of the non-periodic sequencing of gestures that makes use of the notion of optimization to generate biologically plausible fluid movement. The domain employed is that of speech, and we will adopt a more or less conventional understanding of speech production as the progressive attainment of a sequence of individual discrete behavioral goals. Relying on the well-known theory of Articulatory Phonology (Browman and Goldstein, 1990b; Browman and Goldstein, 1992), we understand these discrete goals to be constrictions in the vocal tract achieved by individual gestures. Gestures may be viewed as discrete, context-free goals, in keeping with the requirements of a theory of linguistic contrast. In this way, they may act as constituents within a theory of phonology. As gestures, however, they are simultaneously units of action, expressed in an embodied system, and crucially determined by their physical realization. That is, the characteristics of articulatory movements driven by the gestures are straightforwardly attributable to the physical properties of the vocal tract constituents, like mass, rigid boundaries, etc. This dual interpretation of gestures as units of discrete contrast and as units of action goes some way towards addressing the need for unifying our understanding of how symbol-like systems may be instantiated in an embodied system, without the need for a thoroughly unconstrained translation from one domain to an entirely incommensurable one (Lashley, 1951; Fowler et al., 1981; Harris, 1987).

The notion that optimization principles may be employed in understanding why movement takes one form rather than another is not new. Hogan and Flash (1987) attributed much of the grace of natural movement to minimization of jerk (the first derivative of acceleration), Uno et al. (1989) attributed it to the minimization of torque change. Jordan et al. (1994) sought to account

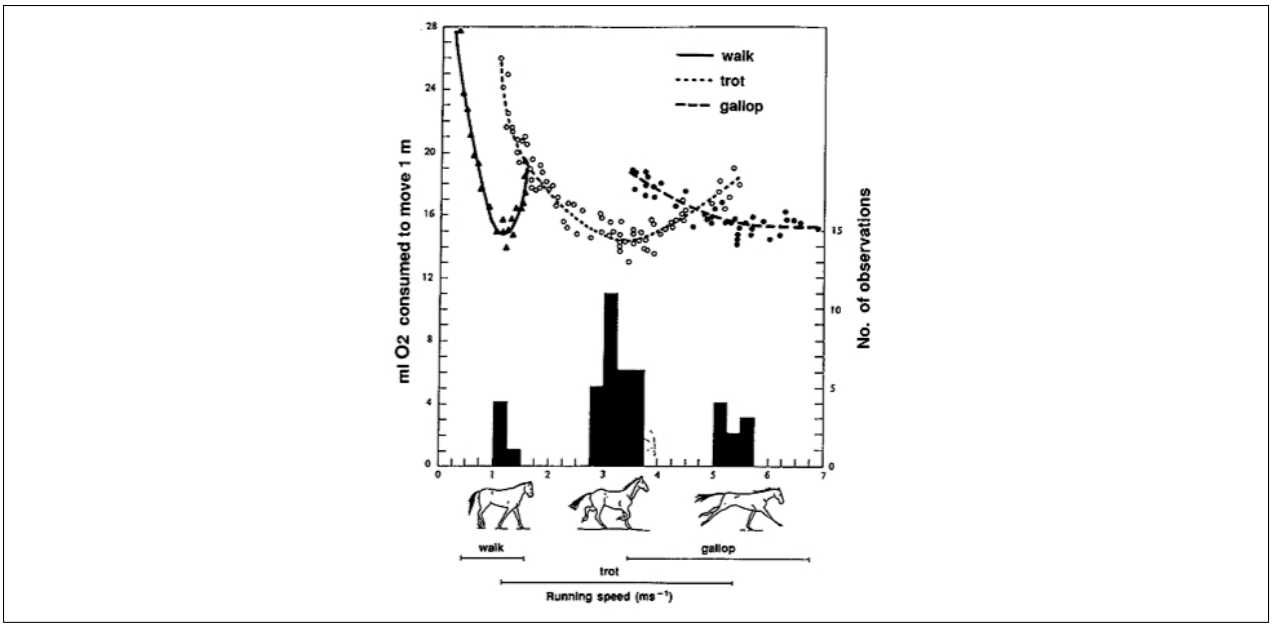

Figure 1: Top: oxygen consumption of a horse on a treadmill running at a range of speeds in three gaits. Bottom: histogram of observed speeds at each of three gaits. Reproduced with permission from Hoyt and Taylor (1981).

for the trajectory of point to point movement by minimizing spatial deviation from a straight line.

A further example of evidence that fluent natural biological movement is in some sense optimal is provided by the work of Hoyt and Taylor (1981) as shown in Fig. 1. The figure shows two data sets. The scatter plot with fitted curves at the top plots the oxygen consumption measured from three remarkably compliant horses running on a treadmill at a range of speeds. For each of the three gaits studied, there is a clear relationship between O_2 consumption and speed, such that a narrow and very well-defined minimum can be found for a walk, a clear but broader minimum for a trot, while the gallop may also have a shallow minimum, but the faster data, at which consumption may rise again, is not probed. The histogram at the bottom of the figure shows counts of observations of the same gaits as a function of rate in the open paddock. For all three gaits, naturalistic observations are only made at speeds corresponding to optima in the oxygen consumption function, suggesting that locomotion is optimal with respect to this one variable at least. Evolution does not work by optimizing a single variable, however, and it is entirely possible that gait selection is influenced by a wide variety of criteria, such as, for example, bone strain (Biewener and Taylor, 1986). In general, if a movement can be shown to be optimal with respect to some criterion, X, that should not exclude the possibility that it is also optimal with respect to Y and Z. Optimization with respect to specific costs may also be interpreted as an operationalization of the concept of efficiency in movement.

The relevance of the notion of efficiency in shaping skilled action has been supported by subsequent research. Anderson and Pandy (2001) used a dynamic optimization approach to show that the requirement of minimum metabolic energy expenditure per unit distance determines many salient features of human gait. The shape of the path taken in reaching movements also seems to be instrumental in minimizing energy expenditure (Nakano et al., 1999).

Efficiency considerations in skilled action underwrite extensive research efforts in motor control and provide a route to understanding learning as being based on optimality principles. For example, the optimal feedback control strategy proposed by Todorov and Jordan (2002) suggests how motor systems might harness the high number of degrees of freedom in reaching and grasping tasks. Optimization allows a simple computer model to discover bipedal control strategies resulting in walking and running (Srinivasan and Ruina, 2006). For an extensive review and discussion of the role of optimality principles in sensorimotor control and learning research, see Todorov (2004) and Wolpert et al. (2001).

There is one important difference between the effortless action of one primate swinging between branches, and a second, hairless one speaking naturally. The nature of the branch-swinging task is manifestly influenced by the morphology of the monkey, but also by the physical properties of the branches themselves, and the opportunity they afford for action. In speaking, the behavioral goals are necessarily formulated with respect to the patterns that can be produced by, and within, a human vocal tract, without reference to any external constraint. Speech thus has emerged as a system of contrastive pattern production, where the elements of contrast have been "found" within the space of potential movement of the system itself. As Lindblom has observed, "(s)ound systems are (in part) adapted to be spoken" (Lindblom, 1999). The study of efficiency and optimality in the production of speech is thus a contribution to our understanding of what speech is.

One of the prime examples of this approach to speech research is found in Dispersion-Vocalization Theory, that motivates an account of the observed distribution of vowels in acoustic and articulatory space (Schwartz et al., 1997). Different languages have very different vowel inventories, but regardless of the number of vowels, their location in a vowel space, that may be defined using either acoustic or articulatory frames of reference, is highly predictable. By minimizing a weighted sum of cost components that take into consideration both stability in production and contrast in perception, i.e. "non-linguistic constraints on possible speech sounds" (Schwartz et al., 1997, p. 256), Dispersion-Vocalization Theory generates distributions that closely match vowel systems naturally occurring in world languages. Vowel distributions can thus be understood as being "optimal" in a precise sense.

The cost functions of Dispersion-Vocalization Theory are not directly linked to metabolic energy expenditure accompanying motor action. In fact, many speech scientists frown upon the idea that efficiency of articulatory motor action—minimizing articulatory effort—plays any role whatsoever in shaping speech phenomena. They argue, for example, that energetic considerations are at best marginal, as the relatively slight masses are acted upon by disproportionately powerful muscles (Keller, 1987), or that the rise of ubiquitous variation in speech patterns is predominantly perceptually driven (Ohala, 1989). According to their view, the rich variety of speech phenomena observed is largely a by-product of disembodied linguistic rules being realized by a somewhat messy physical substratum: the human vocal tract.

On the other hand, there has been a steadily growing move to characterize speech in light of a dynamic, embodied motor control paradigm and to interpret speech phenomena as strictly emergent from a set of non-specific constraints, among which we might include the efficiency of underlying articulatory action. Lindblom (1983) provided a strong argument for the role of articulatory effort in explaining a number of general characteristics of spoken language. In his Hyper- and Hypospeech Theory, inspired by the well known Fitts' law of speed accuracy trade-off for movement control (Fitts, 1954), Lindblom (1990, Lindblom et al. 1995) accounted for the variations accompanying speaking rate changes and environmental influences in terms of trade-offs between perceptual clarity

and production efficiency. Several recent models of speech production, most notably the influential neural network model DIVA (Guenther, 1995; Guenther et al., 1998) employ different control strategies based on efficient realization of individual speech movements to reproduce phenomena associated with speech acquisition, coarticulation and speaking rate changes.

Critics of the articulatory efficiency approach are right in one aspect: in contrast to locomotion research, it is all but impossible to measure directly the metabolic energy expenditure associated with speech production. Speech scientists therefore must rely heavily on models incorporating various accounts of energy expenditure and compare the predictions generated in this way with empirical data. Recent years have seen many studies using kinematic records or the acoustic characteristics of recorded speech (e.g., looking at peak movement velocity) as measures of production efficiency (Perkell et al., 2002; van Son and Pols, 2002; van Son and Pols, 2003). These experiments provide tentative support to the notion of a lawful association between high-level speech phenomena and economy of movement.

1.1. Modeling goals

Our work addresses an important and long-standing challenge: how do we account for the relative *timing* of multiple parts of the body acting in concert in the service of behavioral goals (Lashley, 1951)?

Our proposed solution to this challenge is to take account of both the physical, inertial, properties of the effectors, and the boundary constraints imposed by the behavioral goal, and to demonstrate that these suffice to determine the relative timing of the motion of the coordinated parts. Speech provides an important domain for testing these ideas, but the principles underlying our work are quite general. We choose to implement these ideas within the established modeling framework of Articulatory Phonology and its Task Dynamic implementation, but those elements of our model that provide a principled answer to the stated challenge are independent of both these models.

Using Articulatory Phonology theory as an implementation platform provides us with an explicit means for treating of the timing of the articulatory movements. Articulatory Phonology makes use of a very restricted set of primitives. Gesture on- and offsets are specified, along with a small set of dynamical parameters, such as stiffness, and out of this limited set, highly nuanced movement trajectories are produced. In comparison, most other models of speech production—even those that in principle facilitate such discussion—implicitly or explicitly shy away from the details of timing in sequential action. For example, DIVA "does not address many important issues concerning the control of timing in speech production" (Guenther, 1995, p. 618). Each articulatory movement is triggered at the moment when the previous movement successfully reaches its target. This modeling decision of the DIVA authors is, however, at variance with the long known fact that articulatory movement towards a second vowel in vowel-consonant-vowel (VCV) sequences often starts *before* the target of the intermediate consonant is reached—in fact, depending on context, movement towards the second vowel frequently begins even before the onset of the movement towards the consonant (Öhman, 1966; Löfqvist and Gracco, 1999). It is precisely this type of inter-gestural timing behavior that we believe is best understood to be a consequence of the embodied character of the speech production system, and is most likely a result of the drive for efficiency common to biological systems. Emergent phenomena of this type, along with their conditioned variation, are our principal focus in this paper.

The choice of Articulatory Phonology and its Task Dynamic implementation entails a commitment, at present, to gestural primitives, and to spatial targets in articulator movement. These

assumptions allow the development of means to optimize the movement of the diverse parts of the vocal tract during speaking. There is a body of work that argues quite persuasively for targets that are best defined in acoustic terms (Guenther et al., 1998), and targets that are distributed, rather than punctate (Guenther, 1995). These are essential characteristics of the DIVA model. DIVA seeks to provide a model of real-time movement control, while our goals are rather different, as we seek to account for the observed form of fluent movement, but not to derive it in real-time. Nevertheless, both DIVA and our model combine production and perceptual constraints in the derivation of movement forms, and they thus share some important structural features. The relative merit of acoustic or articulatory targets will be clearer as both models continue to flesh out their respective ambitious agendas.

The ordered set of gestures required to generate an utterance provides a well-defined sequence of behavioral goals. To these, we add some degree of intentional control by defining two dimensions of context-specific variation which can be glossed here as *clarity* and *rate*. These can be seen as the dimensions of intentional or voluntary control at the disposal of a speaker. Collectively, these constitute the boundary constraints that provide half of the solution. The other half comes from the physical properties of the articulators, specifically, from their inertial properties. The articulators within our model have masses that constrain their movements. We do not attempt detailed anatomical modeling, but rather seek to demonstrate how inertial constraints can, in principle, be combined with behavioral constraints to define an optimality function. We then employ computational procedures to identify sequences that are optimal in this well-defined sense.

Importantly, our model is not a real-time production model. The task of identifying optimal sequences is computationally expensive, and necessitates many simplifications within the model. But it has the significant advantage of making strong predictions about the optimal form of movement. In this sense, it is an attempt to provide an initial response to the problem of serial order that Karl Lashley phrased thus:

> This is the essential problem of serial order; the existence of generalized schemata of action which determine the sequence of specific acts, acts which in themselves or in their associations seem to have no temporal valence. (Lashley, 1951, p. 323/324).

1.2. Articulatory Phonology

Articulatory Phonology (Browman and Goldstein, 1990a; Browman and Goldstein, 1992) assumes that primitive actions of the vocal tract articulators, called *gestures*, are the basic atoms out of which the phonological structures of utterances are formed. The gestures are abstract characterizations of coordinated task-directed movements of the articulators within the vocal tract. They form a limited set of presumed basic building blocks that give rise to the complex motion patterns of the vocal tract components. These motion patterns are combined (and sequenced) in order to produce a series of articulatory constellations resulting in an appropriate sequence of acoustic events—an utterance.

These articulatory constellations form characteristic *constrictions* of the vocal tract. Each gesture is a member of a family of functionally equivalent movement patterns of relevant articulators coordinated in order to form and release a particular vocal tract constriction. A speaker producing a bilabial stop /p/ creates an occlusion of the vocal tract by forming (and releasing) a closed labial constriction. She or he moves the lips towards each other until they collide and block the airflow passing through the oral cavity. Simultaneously, the speaker closes the velar aperture and widens

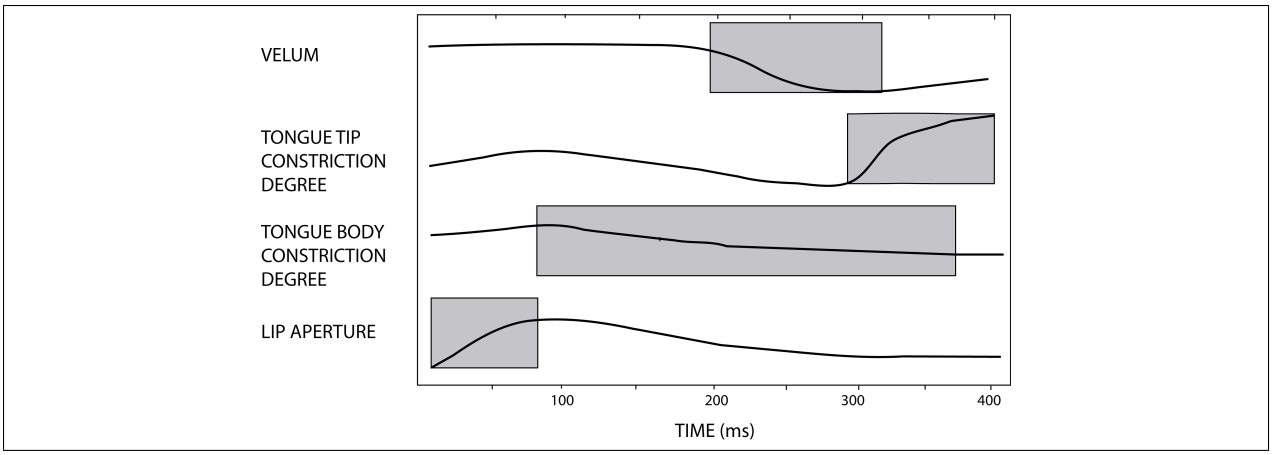

Figure 2: Simple gestural score and associated tract variable trajectories. Adapted with permission from Browman and Goldstein (1995).

the glottis to form a closed velar and wide laryngeal constrictions. These three dynamic gestures involving various articulatory subsystems thus participate in the production of the articulatory and acoustic event traditionally labelled as an unvoiced non-nasalized bilabial stop. As seen in this example, gestures do not correspond to either features or segments, the phonological primitives in traditional theories. Rather, "they sometimes give the appearance of corresponding to features, and sometimes to segments" (Browman and Goldstein, 1992).

Each particular constriction—and, consequently, the gesture involved in its formation—is specified by the descriptors capturing the target position of the relevant end effectors: the *constriction degree* and, in some cases, the *constriction location*. These descriptors correspond roughly to the vertical and horizontal dimensions of the target position. An alveolar stop /d/ is thus characterized by the constriction degree value "closed" (or, numerically, zero) and the constriction location value "alveolar" (numerically expressed as a position of the alveolar ridge in some appropriately chosen coordinate system). Some gestures, for example velum closing, are conceived as uni-dimensional as the constriction location is uniquely determined by their functionality.

At any given time, the level to which these given targets (independently, the degree and the location) are reached is captured in a form of the *vocal tract variables*. Each gesture is associated with one or two tract variables. The target constriction degree and constriction location are, in fact, expressed as appropriately chosen values of the corresponding tract variables. In the case of /d/, for example, these correspond to particular numerical target values of the tongue tip constriction degree (TTCD) and the tongue tip constriction location (TTCL) tract variables.

Each tract variable is associated with a set of model articulators which are involved in the formation of the given constrictions. The labial closing, for example, engages the upper and lower lip. The lips are directly involved in closure formation, their distance determines the degree to which closure has been achieved at any given time—i.e. the value of the lip aperture (LA) tract variable. The jaw, meanwhile, also participates in this task, albeit indirectly: its position in space impacts the absolute position of the lower lip attached to the jaw.

Fig. 2 illustrates a simple, partial, gestural score underlying the production of the phoneme sequence */pan/*. Four tract variables are shown, and several, including tongue tip/body constriction location and glottal gestures have been omitted. The gestural score proper comprises just the filled

rectangles that specify the periods during which individual gestures are active. The agility of the system in attaining the target position prescribed by the active gesture is represented by a *gestural stiffness* parameter. Although gestural score, constriction target positions and corresponding gestural stiffness values fully determine the behavior of the vocal tract, a further mechanism is needed to compute the trajectories of the associated tract variables (shown in the figure as solid lines), and from those, then, the corresponding movements of the model articulators in a synthetic vocal tract (not shown). These additional pieces are provided by the theory of Task Dynamics, originally formulated to model limb movement (Saltzman and Kelso, 1987), and later extended to model speech movement (Saltzman and Munhall, 1989).

1.3. Task Dynamics

Individual gestures are associated with corresponding tract variables. Within the original Task Dynamic model (hereafter, TD), each tract variable is modeled as a simple mass-spring system, and each is, in principle, independent of all other tract variables. Once the gesture is active, as specified in the gestural score, the tract variable moves from an initial position of displacement to a resting equilibrium, or target. This movement is smooth because it is the solution of a second-order dynamical system:

$$\mathbf{M}\ddot{\mathbf{z}} = -\mathbf{K}(\mathbf{z} - \mathbf{z}_0) - \mathbf{B}\dot{\mathbf{z}}. \tag{1}$$

where $\mathbf{M}$, $\mathbf{K}$, and $\mathbf{B}$ are vectors containing masses, stiffnesses and damping coefficients, respectively, while $\mathbf{z}_0$ is the target position of the tract variable $\mathbf{z}$. Not all of the parameters of this set of mass-spring systems are actually used, however. In order to ensure that each system generates movement that progresses directly towards a target position, without undershoot or oscillation, the vector of damping coefficients, $\mathbf{B}$, is set to ensure critical damping, and is thus analytically derivable from $\mathbf{M}$ and $\mathbf{K}$. More importantly, each tract variable has an abstract mass, arbitrarily set to 1. The values in the mass vector $\mathbf{M}$ are thus not linked to the masses of the vocal tract components, and tasks are not meaningfully embodied.

When a gesture is active (as specified in the score), its corresponding equation kicks in and the tract variable moves towards its target. Several tract variables may be simultaneously active, but they are not coupled to one another, and so are independent. The articulators, on the other hand, are yoked together and stand in a many-to-many correspondence with the tract variables. The Tongue Body Constriction Degree tract variable, for example, needs to influence tongue body articulator position, but that in turn is anatomically yoked to the jaw. The jaw, in turn, is affected not just by this one tract variable, but by all tract variables associated with tongue tip position or lip aperture. A mapping is defined between the tract variable space and the model articulator space, as described in Saltzman and Munhall (1989). This mapping ensures that the influence of simultaneously active tract variables on single articulators are appropriately blended together. The net result is the transformation of the gestural score into the real-time motion of model articulators. These can in turn be used as input to a simple articulatory synthesis routine that generates sound, although for most purposes, it is the fine detail of movement that is of primary interest (Rubin et al., 1981).

The articulatory phonology approach, together with its task dynamic implementation, have been very influential. They have provided new and compelling insights into processes such as the apparent deletion of segments that may arise from gestural overlap, the appearance of epenthetic

vowels, assimilation, and a range of other phenomena (Browman and Goldstein, 1990b). Much of the fine detail of the resulting articulator movement depends critically on the timing details specified in the gestural score, and on the remaining free dynamical parameters, the stiffness coefficients (**K** in Equation 1).

Initially, gestural scores were drawn by hand. That is, in order to produce a given sequence such as */pan/*, trial and error was used to obtain a score and associated stiffnesses that would generate suitable movements. There were attempts to learn appropriate settings using recurrent neural networks trained on articulatory data (Saltzman and Munhall, 1989), but these were not very successful, perhaps due in part to the variability that is to be expected in skilled movement data. Just as handwriting and gaits speak not only of abstract goals, they also reflect the highly individual solution adopted after much experience by a specific individual. The possibility that the relative timing among gestures was relatively invariant, that is that a consonant would reliably begin at an invariant phase of the tauto-syllabic vowel, for example, proved too restrictive (Browman and Goldstein, 1990b). The idea of highly constrained mutual timing relations among gestures was further developed with the introduction of phase windows, an approach that suggested that the set of phase relations (i.e. relative timing) among gestures was multiply determined, with each factor contributing independently to a probabilistic distribution of phase values (Byrd, 1996). Factors could be derived from linguistic factors, but could also reflect extra-linguistic system-specific properties.

More recently, the task dynamic architecture has been extended to include planning oscillators (Nam and Saltzman, 2003). These are abstract timing oscillators at a further remove from the physical properties of the articulators. Each gesture is associated with one planning oscillator, and coupling among the oscillators provides the stable patterning observed in time. Furthermore, planning oscillators may be posited at multiple levels of the prosodic hierarchy, allowing the incorporation of rhythmic and phrasal constraints on timing.

In recent work, we have suggested an alternative approach to deriving appropriate timing (and stiffnesses) for gestural scores. To do this, it was necessary to revise the task dynamic model substantially, making the behavioral goals or tasks embodied. The Embodied TD model has been introduced in Simko and Cummins (2010), based on Simko (2009), and full details are provided in either of these. In the following section we highlight the aspects of this approach that are relevant in the context of the presented work.

2. Embodied Task Dynamics

In the initial implementation of the Embodied Task Dynamic model, we make use of a greatly simplified model articulatory system, as illustrated on the left of Fig. 3. Only the vertical dimension of jaw, lip, and tongue movement are represented within the model. The tongue body, which is attached to the jaw, allows contrasting tongue body positions for the vowels /i/ and /a/. The tongue tip is attached to the tongue body. The lower lip also attaches to the jaw, while the upper lip has a fixed point of support. At present, no glottal or velar modeling is done. Along with the two vowels, this simple vocal tract can produce a bilabial and an alveolar stop, which we identify as /b/ and /d/, respectively, although, in the absence of a glottal component, a voicing contrast is not represented. This very minimal architecture is deliberately chosen so that we can focus on the principles underlying the timing and sequencing of the movement of embodied components.

As we are concerned with describing a multi-component system in which the components syn-

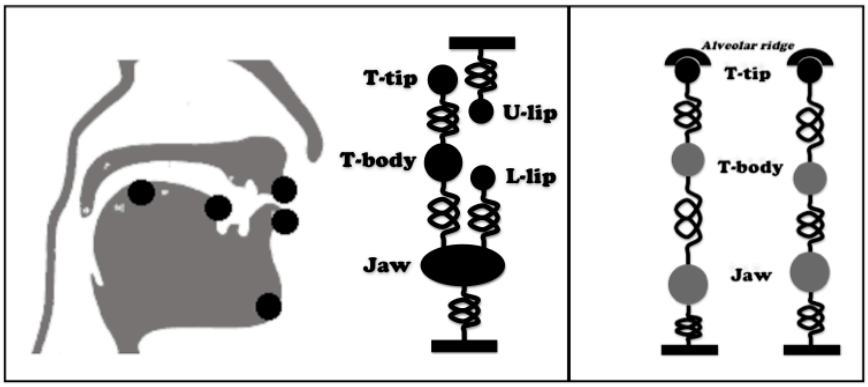

Figure 3: (Left panel) Jaw, tongue body, tongue tip and lips, as they relate to a human vocal tract (left), and as implemented (right). (Right panel) Two behaviorally equivalent articulator configurations are shown.

ergistically cooperate in the achievement of high level behavioral goals (gestures), we need to be careful to describe the behavior of the components within appropriate frames of reference. When we discuss tract variables and constriction targets, we are interested in capturing the collective behavior of multiple components, and we do so by expressing quantitatively the position of the relevant end point(s) in a vocal tract-based coordinate system. These functionally relevant end-points, we call *end effectors*. Our numerous simplifications ensure that, at present, a single constriction target is associated with a single tract variable, which in turn is realized by the movement of a single end effector. The sole exception to this lies in the lips, where the tract variable of Lip Aperture (LA) depends on the position of two end effectors, the position of the upper and lower lips.

A given end effector position may, in general, be achieved in many ways. When the jaw, tongue body and tongue tip cooperate in achieving an alveolar closure, for example, the goal-based end effector position (tongue tip at the alveolar ridge) does not uniquely determine the position of the individual components, as shown in the right hand panel of Fig. 3. This kind of trading off between individual articulators in the service of a single high-level goal is typical of coordinative structures, as documented for example by Bernstein (1967), or in perturbation studies (Abbs and Gracco, 1983; Kay et al., 1991).

We will also need to consider the component articulators individually. This is essential when we come to calculating articulatory effort. The relative position of the tongue tip can be expressed with reference to a coordinate system centered at the tongue body. In this way, the tip movement and associated effort can be evaluated independently of jaw or tongue body movement. Tongue body position can be expressed relative to the jaw, while the jaw itself moves relative to the vocal tract coordinate system. The upper lip has a fixed support, but the lower lip likewise can be considered together with the jaw, as an end effector, or relative to the jaw, as an independent component with its own mass and stiffness. When we are considering the relative position of articulators in this fashion, we will refer to them as *pure articulators*, in order to emphasize their consideration in isolation.

The thoroughly embodied character of our model demands that individual articulators have individual masses and stiffnesses. Thus, although the task-level of description is best done with

reference to end effectors in a vocal tract coordinate system, the computations done over the components will be carried out over the pure articulators. Conceptually, pure articulators are the masses upon which forces act directly.

In framing our model, we make use of a relationship between gestural activation specifications (the score) and vocal tract response (the gestures) that is closely related to the established Task Dynamic implementation of Articulatory Phonology (Saltzman and Munhall, 1989). There are, however, two major differences in the way the response of the vocal tract model to the gestural activity patterns is resolved. Both of these changes arise because we ensure that the target-oriented action of the vocal tract is meaningfully embodied. By this we mean that the system's behavior is crucially influenced (1) by the *masses* of articulators and (2) by the *physical boundaries* participating in realization of some of the gestures. We will now address each of these in turn.

2.1. Articulators have masses

As in the traditional TD model, the task dynamics of the system is defined at the level of tract variables. The active tract variables are modeled as a mass-spring dynamical system (see Eq. 1) with the equilibrium points representing the target positions and the stiffness parameters quantifying the agility of the system's response to currently active gestures. Unlike the classic TD model, the mass parameters, however, are not arbitrarily set to 1, but instead are interpreted in terms of physical masses influencing the movement of pure articulators. In this way, the inertial properties of the articulators become constraints on movement that can help to selectively identify some movements as more efficient than others.

Technically, the introduction of non-unit masses for articulators requires an alteration to the original method of resolving the redundancy of the pure-articulator-to-tract-variable mapping using a scaled pseudo-inversion. Less technically, the representation of mass at the level of the individual pure articulator allows these components, the pure articulators, to function as embodied degrees of freedom of the vocal tract. Like in the original TD implementation of Articulatory Phonology, the active tasks give rise to coordinative structures: pure articulators act in synergy to achieve the required goals defined at the level of tract variables. However, our method of inverting of the redundant pure-articulator-to-tract-variable mapping guarantees that the resulting movements of the individual pure articulator arise from the lawful action of forces upon its mass (Simko and Cummins, 2010).

The gestural stiffness determining the swiftness with which the end effectors move towards the prescribed targets is distributed among the pure articulators that themselves act as coupled critically damped mass-spring dynamical systems. Importantly, we specify *relative gestural stiffness* only: vocalic gestures act proportionally less swiftly than rapid consonantal ones. The set of stiffness parameters as a whole is scaled by a single overall system stiffness scaling parameter, $\mathbf{k}$. This system parameter is not fixed, but emerges from the optimization procedure, which seeks to optimize gestural onsets, offsets, and $\mathbf{k}$ simultaneously.

Within conventional TD, the dynamical systems associated with individual tract variables are mutually independent, or uncoupled. Scaling the task dynamics with respect to the underlying physical parameters of mass and stiffness, as we do, has the consequence that the differential equations desribing the motion of tract variables are no longer uncoupled. They exert reciprocal influence upon one another. Formally, the matrix $\mathbf{M}$ in Equation 1 is not necessarily diagonal. This constitutes a significant departure from the basic principles of traditional task dynamics and Articulatory Phonology. A substantive insight expressed within our model is that gestures are nowhere

completely context-free. Their realization depends in part upon anatomical and physiological links between the articulators employed by concurrently active gestures.

2.2. Physical boundaries matter

The vocal tract articulators move within the physical boundaries of the oral cavity. These boundaries are instrumental in the production of stop consonants as modeled in this work. The alveolar closure, for example, is achieved when the soft tissue of the tongue tip collides with a part of hard palate, the alveolar ridge. Unlike traditional TD, we explicitly model these boundaries using a non-linear damping component influencing end effector dynamics. When relevant end effectors are sufficiently close to each other (lips) or to the oral cavity boundary, their movement is halted by the damping force which increases non-linearily with decreasing distance. These collisions are used in evaluation of the successful realization of stop consonants.

2.3. Mapping between coordinate systems

As in TD, the details of articulatory action are jointly determined by having two related levels of description—i.e. coordinate systems—each imposing appropriate constraints pertinent to the objects expressible within the given coordinates. The tract variable coordinate system, by virtue of representing task-relevant goals, ensures that the model articulators are constrained to act in synergy, and that the end effectors achieve the constrictions prescribed by the gestural score. The physical oral cavity boundaries used to model the collisions that provide articulatory closure for stop consonants are also represented as unmoving objects in this coordinate system. The physical attributes of the vocal tract articulators, the degrees of freedom and the associated masses acted upon by the muscles, are captured in the second system: the pure articulator coordinate system, in which each articulator is described as a point with reference to its nearest articulatory anchor (tongue tip anchored to body, body to jaw, jaw position relative to jaw hinge, etc).

A somewhat simplified account of the relationship between these two coordinate systems is as follows: The complete set of pure articulator coordinates completely suffices to fix a point in the tract variable system. In fact, this mapping is redundant, with many articulator constellations potentially corresponding to the same point in tract variable space, as shown in Fig. 3, right panel. A non-invertible matrix thus specifies this mapping. The dynamics unfold in tract variable space, however, so we need to map back from that coordinate system into the space of articulators. This is done using the technique of pseudo-inversion, to provide a mapping that is, in a precise sense, optimal. These two mappings introduce constraints in both directions between the coordinate systems. The task dependent coupling among pure articulators that is thereby introduced guarantees that the vocal tract components act in synergy in order to achieve given goals. At variance with the classic approach, however, the pure articulator action is also constrained by the physical properties of the modeled anatomical structure: each pure articulator behaves as a damped mass-spring system acting on an appropriate mass in a way analogous to the real muscle structures acting on the physical masses of the vocal tract articulators. This constraint imposes additional dynamical coupling at the tract variable level—the tract variables are no longer uncoupled, and their behavior is no longer independent of the embodied nature of the underlying anatomical structure.

As full details of the embodied TD model have been presented before (Simko and Cummins, 2010), we omit much of the complexity here to focus instead on the principal benefit derived from the new model. Because we can now quantify articulatory effort, it is possible to posit a parametric

cost function that will, in turn, allow the definition of an optimal sequencing of gestures. It is to this that we now turn.

3. A Parametric cost function

Before detailing the components of our parametric cost function, it might be timely to look ahead at the distal goal of the exercise. Fig. 4 illustrates 10 steps in a protracted search for an optimal gestural score for generating the sequence */ibad/*. At the top is the starting score, in which the four gestures are arranged in sequence, without overlap, and a system-wide stiffness value, used in scaling all individual stiffnesses within the system, is set to a high initial value. Based on a parametric cost function, to be described below, this configuration can be assigned a cost C. Now, through variation of the onset and offset times of the individual gestures, and of the system-wide stiffness coefficient, $\mathbf{k}$, it is possible to implement a simple gradient descent procedure to arrive at a more efficient score, with a lower value of C. Several steps in this procedure are illustrated, ending in the final configuration which resists improvement despite continued search. This final configuration is then taken to be optimal, and the embodied task dynamic model then generates the associated articulator movements for this sequence. Given the initial desired sequence of gestures, both the optimization procedure and the subsequent conversion to articulatory trajectories are fully automatic. The resulting trajectories can then be assessed for plausibility (in the first instance) and can be compared to articulatory data.

The cost function we propose is a provisional attempt to capture some high-level constraints known to influence gestural precision and timing. The cost function has three independent components which combine in a weighted sum:

$$C = \alpha_E E + \alpha_P P + \alpha_D D,$$

where α_E, α_P and α_D are simple scalar weight coefficients. E measures *articulatory effort*, which is derived from the overall force (acting on pure articulator masses) involved in utterance production. P is a *parsing cost*, inspired by the desire to include a perceptual measure related to communicative efficiency, and D expresses the relative importance of utterance duration. The three components provide a set of interacting constraints on an utterance, which collectively determine the optimal sequence of gestures and system stiffness. We will now discuss each in turn. In describing how each component is computed, it will be necessary to delve into detail, much of which is contingent and specific to our model. The reader for whom this is of secondary importance may freely skip to Section .

3.1. Articulatory effort

Movement is costly. The first component of our composite cost function, E is a measure of physical effort required to produce an utterance. It is minimized by doing nothing, but of course, this tendency to be lazy is offset by component P, which penalizes communicative difficulty.

The articulatory effort cost component is linked to the elusive concept of articulatory ease that has proven notoriously difficult to quantify during analysis of experimentally recorded speech data. Various approximations, most notably peak velocity of articulator movement, are traditionally used in data analysis (Perkell et al., 2002). Our modeling platform, however, offers a direct and simple means to evaluate physical measures linked to economy of effort.

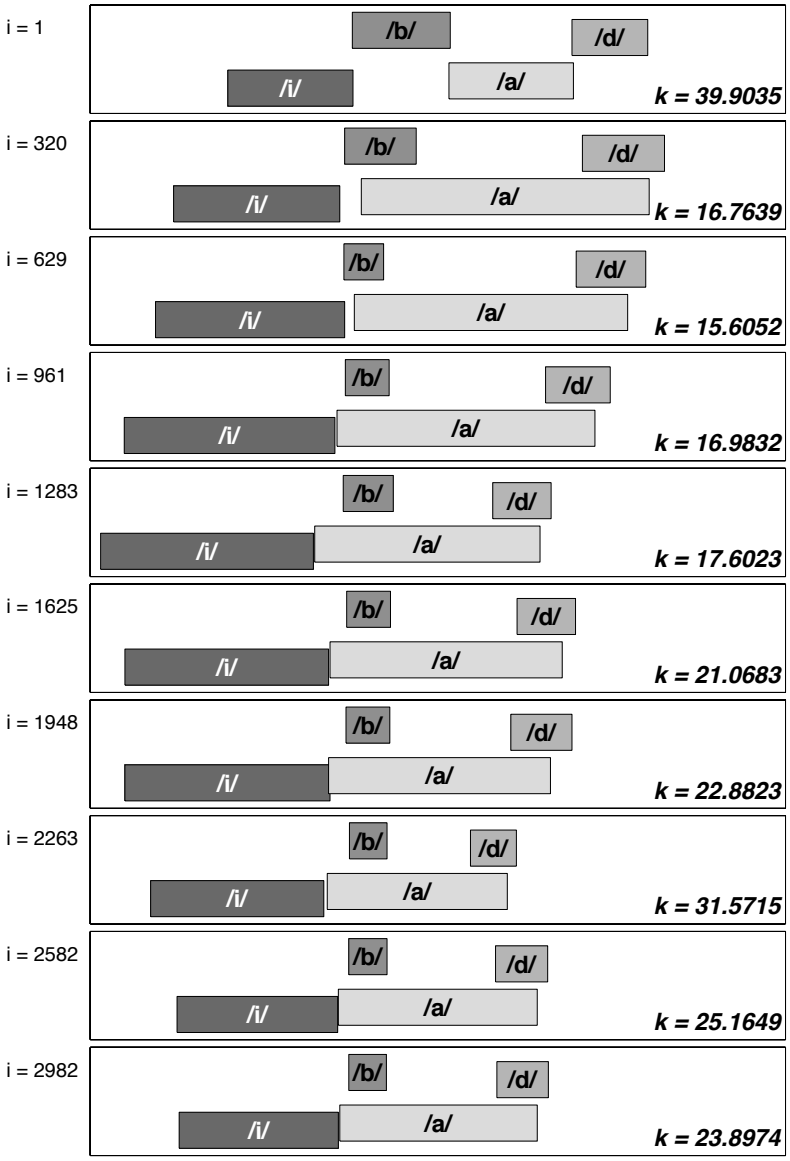

Figure 4: Sequence of gestural scores and corresponding values of overall stiffness parameter evaluated during a search for an optimal realization of sequence */ibad/*. Number i indicates the number of steps taken. Overall cost decreases monotonically from top to bottom.

As argued above, the pure articulators of our model represent the constituents of the human vocal tract whose dynamic characteristics are directly linked to the underlying muscle structures involved in utterance realization. For example, the tongue tip pure articulator is the position of the tongue tip relative to its proximal anatomical attachment, the tongue body. The tongue body pure articulator is then the position of the tongue body with respect to the jaw to which it is anatomically linked, etc. In order to produce the movement of the tongue tip end effector towards the alveolar ridge, the vocal tract muscles act in synergy: the muscles linking the tongue tip to the tongue body act on relatively slender mass of the tongue tip alone, the muscles linking the tongue body to the jaw act on the comparatively heavier mass of the entire tongue, and the muscles moving the jaw act on a heavy load including all anatomical structures attached to the jaw (the tongue, the lower lip, teeth, etc).

In Section 2, we said that the computations performed by the model are done at the level of the pure articulators. The solution of the system of differential equations yields the pure articulator positions in time represented by variables y_{UL}, y_{LL}, y_J, etc. The embodied version of task dynamics that drives the pure articulators of the model not only generates appropriate pure articulator accelerations giving rise to coordinative structures realizing a given set of active tasks, but it also guarantees that these accelerations lawfully reflect the masses acted upon by the underlying muscular structures. The forces applied to the pure articulator components represent the actual muscle forces behind the sequential attainment of given speech targets. If, for example, $\ddot{y}_{\mathrm{TT}}$ is the acceleration of the tongue tip pure articulator at a given moment, and m_{TT} is the tongue tip mass, $\vec{F}_{\mathrm{TT}} = m_{\mathrm{TT}}\,\ddot{y}_{\mathrm{TT}}$ is the current force acting on the tongue tip anatomical sub-component during the alveolar closure action. The alveolar closure, of course, also elicits similarly evaluated forces $\vec{F}_{\mathrm{TB}}$ and $\vec{F}_{\mathrm{J}}$ acting on the tongue mass and the jaw, respectively.

We presume that articulatory effort is directly linked to the magnitudes of all such forces driving the anatomical components of the vocal tract during the time course of an utterance realization. Therefore, we evaluate the articulatory effort cost component function as

$$E = \sum_{a} \left(\int_{t_b}^{t_e} |\vec{F}_a| dt \right), \tag{2}$$

where t_b and t_e are the onset time of the first and the offset time of the last gesture in the sequence, the index a ranges over all of the system's pure articulators and $|\vec{F}|$ is the magnitude of the force $\vec{F}$.

3.2. Parsing cost

The articulatory effort cost function defined above represents the production oriented factors of speech production. We now introduce a competing cost term that is related to articulatory precision and clarity of realization of individual gestures in a given utterance. The parsing cost is linked to the demands imposed on the speaker to produce an utterance parseable by the listener in a given situation. The greater the effort imposed on the listener to parse the utterance, the higher the parsing cost. As with the other two components, we can envisage many ways in which a perceptual constraint, or parsing cost could be implemented. The somewhat detailed account that follows is simply one that works within our model.

We presume that the parsing cost is directly and straightforwardly related to the quality of articulatory output of our production model. Therefore, we present here a method of quantitative

evaluation of the articulatory output, which does not take into account the complex, non-linear nature of the relationship between speech articulation and its acoustic counterpart.

In order to assess the articulatory quality relevant to the listener, we consider two aspects of the realization of each individual gesture in an utterance.

The first is an estimate of the precision with which the gestural target associated with an active gesture is achieved. Precision here is a quantity that is inversely proportional to the distance of the tract variable from the target. The more precise the articulation, the lower the cost of evaluating the gesture by the perceiver. We refer to this measure of precision as the *precision estimate of realization* of a given gesture. This precision estimation function is thus an (inverse) measure of articulatory undershoot associated with the given gesture (Lindblom, 1983; Lindblom, 1990).

The second aspect of the production quality evaluation captures the temporal dimension of gestural realization. Again, we presume that the longer the articulatory event associated with the gesture persists, the easier it is for the listener to identify the gesture (Gray, 1942). This measure of gesture production is quantified as a *temporal estimate of realization*. The gestural precision estimate and the temporal estimate of realization are then combined into a single scalar quantity, the realization degree. We now address each of these sub-parts of the parsing cost in detail.

3.2.1. Gestural precision estimate

We presume that the demands posed on the listener are related to the precision with which articulatory targets associated with each sequenced gesture of an utterance are reached.

For a gesture g, the precision of its realization increases as the distance of the tract variable z from the given constriction target z'_g decreases. If z_0 is the value of the tract variable when the system is in its resting, or speech ready, state, we formally define this precision estimate as

$$p_g(t) = 1 - \left| \frac{z'_g - z(t)}{z'_g - z_0} \right|. \tag{3}$$

For each gesture g, the estimate p_g is thus a time function depicting the level of achievement of the gesture's target. Note, that the definition yields a meaningful value for every gesture defined in the model at all times. To avoid the dependence of the precision estimate on anatomical details of the model vocal tract, the function p_g is normalized with respect to the distance $|z'_g - z_0|$ of the realization target from a neutral, or speech ready, position of the tract variable. This ensures that the closer targets are not evaluated as inherently more precisely realized than the farther ones.

The precision estimate function p_g reaches the maximal possible value 1 when the tract variable participating in realization of the gesture g reaches its realization target z'_g; otherwise the value is less then 1. The value 1 thus represents a realization of the given gesture with no articulatory undershoot.

At present, we treat consonants and vowels differently in estimating precision. While vowels may undershoot, consonants are *required* to achieve closure (i.e. $p_g = 1$), or else a punitive cost is incurred that ensures the resulting form will not be deemed optimal. This is somewhat draconian, and a relaxation of this requirement to include principled lenition processes and speech error production is possible in the future.

We refer to any continuous time interval during which the precision estimate value of a gesture is higher than an appropriate threshold as the *realization interval* of the given gesture. The realization interval of a (stop) consonant coincides with the consonantal closure (see below) and the realization

interval of a vowel is defined as the time interval during which the corresponding precision estimate is greater than a realization threshold set arbitrarily to 0.6.

It may happen that the realization intervals for a stop and a vowel overlap. In this event, we recognize that for the listener, the vowel realization will be occluded by the stop closure. Only upon stop release will the otherwise prominent vowel be potentially perceived. That interval during which a gesture is not occluded is its *prominence interval*. In the interests of simplicity, we presume that at any given time there is precisely one speech segment perceived by a listener. In the case of multiple realization intervals overlapping, one of the realized gestures is deemed to be the most prominent, and the overlapped portion is considered a part of its prominence interval. It is important that the prominence interval of a segment be sufficiently long if it is to be perceived.

3.2.2. Temporal realization estimate

The requirement that a gesture be of "sufficient duration" in order to be perceived is inexact. We do not claim that there is a fixed durational threshold delimiting "good" and "bad" realization. Moreover, while it is true that the longer the interval, the easier it would be to identify the realized gesture, we do not presume that the listener's ability to identify the presence of a gesture grows linearly with the interval duration. Rather, we presume that it increases dramatically within a few first tens of milliseconds after the onset of the gesture's prominence interval, and then remains virtually unaffected (Gray, 1942).

We model this durational requirement using a monotonically increasing time function with range $[0, 1)$ triggered (reset to 0) at every onset of a prominence interval of any gesture in the sequence. During the prominence interval of gesture g, the temporal estimate of the realization of the gesture g is thus formally defined as

$$d_g(t) = \frac{2}{\pi} \arctan(c(t - t_1)), \tag{4}$$

where t_1 is the onset time of the prominence interval of gesture g, and c is an adjustment constant. The adjustment constant influences the slope of the temporal estimate function: the higher it is, the steeper the function d_g. This constant accounts for assumed differences between the durational requirements posed on consonants and vowels. In our model, the adjustment consonant is set to a higher level for consonants than for vowels, i.e., the function $d_g(t)$ rises faster for the consonantal gestures than for the vocalic ones. An elaboration of the role of this constant might later be possible to capture both vowel length and gemination phenomena within specific languages.

3.2.3. Realization degree

We now combine the precision estimate and the realization function together. This is done in different ways for vowels and consonants. Taking consonants first, we define the *realization degree* of consonantal gesture g as dependent solely on the temporal estimate function associated with the gesture during its prominence interval:

$$r_g = \max_{t \in [t_1, t_2]} d_g(t), \tag{5}$$

where t_1 and t_2 are the boundaries of the prominence interval (closure) of the gesture g. As the temporal estimate function is increasing, in fact

$$r_g = d_g(t_2).$$

Unlike the temporal estimate and precision estimate functions, the realization degree r_g is a single number from the interval $(0, 1)$ evaluating the perceptual quality of the given segment realized by the gesture g. The higher the realization degree, the easier it is for the listener to identify the gesture in the sequence.

For vowels, we combine the precision estimate with the temporal realization estimate to generate the overall realization degree, so that

$$r_g = \left[\max_{t\in[t_1,t_2]} p_g(t)\right] . \left[\max_{t\in[t_1,t_2]} d_g(t)\right] = d_g(t_2) \max_{t\in[t_1,t_2]} p_g(t), \tag{6}$$

where t_1 and t_2 are the boundaries of gestures g's prominence interval.

The realization degree of consonants and vowels is a quantitative measure of their articulatory quality achieved during the production of an utterance. We presume that this quality is proportional to the cost of parsing the utterance by a listener. The higher the realization degrees of gestures in the sequence, the easier it is to parse the utterance, and, consequently, the lower the associated parsing cost.

The parsing cost associated with processing a single gesture with realization degree r_g thus can be expressed as $1 - r_g$. The overall parsing cost of a sequence $/g_1, g_2, \ldots, g_n/$ of *realized* gestures is therefore defined as

$$P = \sum_{i=1,\ldots,n} (1 - r_{g_i}). \tag{7}$$

The value P is always positive, the lower bound of the parsing cost is 0. This minimum can, however, never be reached in practise: due to the durational element (temporal estimate function), the realization degree of every gesture is always less then 1.

The parsing cost defined this way is, as intended, a measure of overall articulatory undershoot and temporal shortening of realized speech gestures.

When searching for an optimal activation pattern and an overall stiffness value we limit our search to trials realizing a *prescribed*, non-empty sequence of gestures. When an input (gestural score and overall stiffness) fails to produce the required sequence, we assign to it a very high arbitrary precision cost exceeding any value possibly obtainable by Formula 7. This ensures that the optimization procedure remains within the input regions realizing the required sequences, in effect guaranteeing that the computed optimal activation pattern and overall stiffness value together produce the given utterance.

3.3. Duration

The articulatory effort and the parsing cost react in opposite ways to variations in the gestural timing and the system's dynamic parameter **k**. An increase in the system's overall stiffness leads to an increase of the articulatory effort, but to smaller undershoot, i.e. a decrease of the parsing cost. Shortening the activation intervals of individual gestures results in a decrease of the articulatory effort required for utterance's realization, but brings about an increase of articulatory undershoot—an increase of the parsing cost.

As mentioned earlier, this trade off between the production and perception constraints has been conceptualized by Lindblom (1990) in his Hyper- and Hypospeech Theory (H&H Theory) of phonetic variation. A speaker's natural tendency to minimize the articulatory effort leads to an increase of perceptual parsing cost, i.e. less precise and shorter realization of gestures—or

hypospeech. If, however, circumstances require better intelligibility, i.e., clearer speech, the speaker may shift his attention to the listener-oriented parsing cost, and hyperarticulate his utterances at expense of the articulatory effort exerted.

If we were to use only P and E in our cost function, we would have a first pass implementation of H&H theory, and the ratio of the two cost components could serve as an index along the hypo- to hyper-speech continuum. However, as Lindblom himself noted:

> (...) the assumption about H&H variation being one-dimensional is a deliberate simplification which is likely to be revised in the course of further work. (Lindblom, 1990)

One way in which this simplification is obvious is in the interaction between hypo- (or hyper-) articulation and speech rate. H&H theory, without further extension, predicts a rather simple interaction between degree of articulation and speech rate: hyper-articulation should take more time, and hypo-articulated speech should be more rapid. However rate appears to be somewhat independent of this dimension of variation. Gay (1981), for example, reported that changes in speaking rate do not necessarily lead to the consequences implied by H&H Theory. People can speak quickly without undershooting articulatory targets, and they can speak slowly with imprecisely realized underlying gestures. In addition to adjustments of segmental duration and articulatory displacement, the changes in speaking rate can be elicited by means of non-linear alterations of articulatory velocity and intrasyllabic coarticulation. In terms applicable to our modelling platform, the speaking rate can be increased by a non-linear scaling of gestural sequencing patterns and an accompanying adjustment of the model's dynamic parameter, the system stiffness.

The obvious measure indicating changes in speaking rate is the duration of produced utterances. Therefore we include the duration of the realized gestural sequence as another cost component used for evaluation of the overall associated with utterance production.

The definition of the *duration cost* D associated with an utterance production is straightforward: D is the length of the time interval starting at the onset of the activation interval of the first active gesture in the utterance's gestural score and ending with the offset of the last gesture.

Unlike articulatory effort and parsing cost, the duration cost is not directly associated with any expenditure of energy on behalf of either speaker or listener. Rather, it represents a global constraint imposed on the manner of speech production reflecting an intentional choice of the speaker with respect to speaking rate.

One way of regarding the parametric cost function, then, is as a representation of the low-dimensional space of intentional control available to a speaker. While a speaker might be readily able to speak more or less clearly, or more or less rapidly, each of these dimensions of variation gives rise to a host of empirical changes to manifest speech patterns that are not individually controlled or controllable by the speaker.

3.4. Optimization

Faced with the requirement of finding the input constellation (gestural score and overall system stiffness) that is optimal with respect to the three cost functions just defined, we are presented with a multi-objective optimization problem. We approach this problem in a standard way, and define the *overall cost* function C as a weighted sum of component cost measures:

$$C = E + \alpha_P P + \alpha_D D, \tag{8}$$

	E	P	D
stiffness ↗	↗	↘	↘
activation lengths ↗	↗	↘	↗

Table 1: The direction of the covariation between the constituent cost components (E, P, D) and the system stiffness and gestural activation intervals.

where α_P and α_D are the parsing and duration weight coefficients, respectively. (For simplicity, we can scale the coefficients so that the corresponding weight $\alpha_E = 1$). The value of the coefficient α_P determines the position of the solution of this optimization problem on the H&H scale; the coefficient α_D influences its position along the slow-fast speech dimension. We thus explore a two-dimensinal space of articulatory patterning within the model.

The following table presents the influences of some high level properties of the gestural score and system stiffness constellation on the constituent cost functions. It shows, e.g. that a relatively high value of the system stiffness, k, would be associated with a greater degree of estimated effort, and with smaller estimates of parsing and duration cost components.

In order to find a gestural score that is optimal, we begin with a desired sequence of gestures. The simplicity of our vocal tract means that gestures stand in one-to-one correspondence with 'phonemes' and so we can take the sequence */ibad/* as an illustrative example (See Fig. 4). The constraint that any viable sequence must actually realize this sequence of four gestures, that is each must be realized in the correct order, serves as an initial constraint on the search procedure.

To find the optimal input constellations realizing a given sequence we use a simple method related to simulated annealing and implemented in MATLAB. Our implementation of simulated annealing uses a gradient descent optimization method to find local optima of the objective function C. Due to the high complexity of the searched space, the constellations found in this way are then perturbed (replaced by a random nearby constellation) in a simulated annealing fashion and the gradient descent search continues. The parameters that are adjusted at each time step are the onset and offset times of each gesture and the overall system stiffness (Fig. 5, left). All other parameters are fixed.

To find a local minimum of the function using gradient descent, at each step the constellation is adjusted in the direction of the negative of the gradient of the objective function at the current point. The gradient is a vector in the input constellation space which points in the direction of the greatest rate of increase of the objective function C.

The overall cost associated with the production thus decreases with each step, until it "gets stuck" in a local minimum (Fig. 5, right). The perturbation then "releases" the optimization process from this local attractor and the search continues. The search is terminated when a given number of the random perturbations (decreasing in magnitude) fails to lead the optimization procedure to a new local minimum. The constellation reached is then used as a sufficiently reliable estimate of the global minimum of the objective function C. The resulting gestural score, together with the value for the overall system stiffness, minimizes the cost of sequence production.

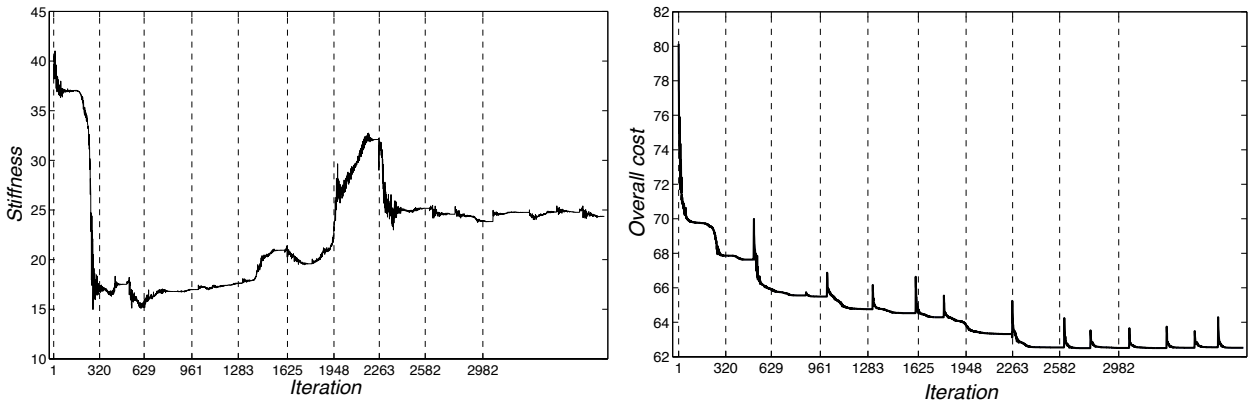

Figure 5: Left: Values of the overall stiffness parameter as a function of steps taken during the optimization of gestural sequence /ibad/. Right: Evolution of the value of the overall objective cost function. See also Fig. 4.

4. VCV simulations

The vocal tract represented within our model is highly simplified. This is a deliberate choice allowing us to focus on the development of a principled procedure for determining the free parameters of a gestural score that generate appropriately detailed movement. Once those procedures are in place, the model can be extended to capture anatomical details of a more realistic vocal tract.

We present here an exploration of the space of parametric variation for the four possible V_1CV_2 sequences where $V_1 \neq V_2$, i.e. /iba/, /abi/, /ida/ and /adi/. In generating these, the parameter α_P ranges over the values {1, 2, 4, 8, 16} and α_D over the values {4, 8, 16, 32, 64}, providing 25 different sequences. For each simulation, we start with a naïve initial gestural score in which gestures are sequenced with no overlap. We set initial overall stiffness to 100 Nm^{-1}. Optimization is carried out as described above. This is done twice for each pair of parameter values, and the better of the two simulations is taken as approximately optimal.

Fig. 6 shows two representative results obtained for the sequence /abi/ with $\alpha_P = 1, \alpha_D = 8$ (left panel) and $\alpha_P = 8, \alpha_D = 8$ (right panel). For a fixed duration cost, these are two quite distinct values of the parsing cost, resulting in disyllables of approximately 220 and 500 ms, respectively. In the figure the top panel shows the final gestural score. The rectangular boxes represent periods of gestural activation, that is periods during which the end effectors are actively attracted towards their respective target positions. In both cases, it can clearly be seen that the consonant activation completely overlaps the continuous vowel activations, thus providing a separation of vowel and consonant tiers, as demanded by most current phonological approaches (Browman and Goldstein, 1990b). In some simulations, there is a small gap between the activation intervals for the first and second vowels, as in the left panel of Fig. 6. Elsewhere, vowel activation is perfectly continuous. It can also be seen that these two simulations generate slightly different sequential orderings of key events.

On the right, the onset of the consonantal activation precedes the inter-vocalic switch by approximately 35 ms, while on the left, it follows it by about 5 ms. We will return to this detail below. The lower panels in Fig. 6 provide kinematic traces for movements of the jaw (thick solid line),

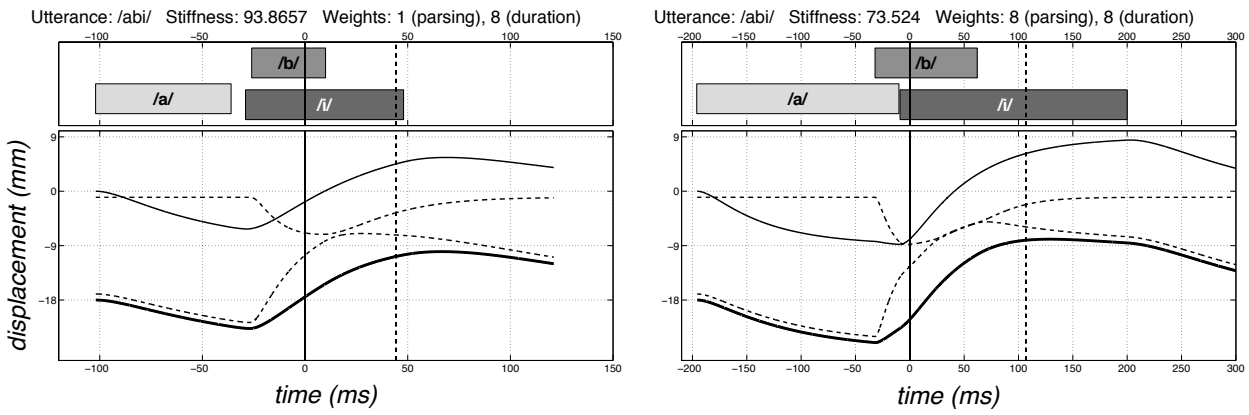

Figure 6: Optimal gestural scores and associated kinematic traces plots for the utterance /abi/. Left: $\alpha_P = 1, \alpha_D = 8$. Right: $\alpha_P = 8, \alpha_D = 8$. Vertical lines mark consonant closure (solid) and release (dashed). Heavy solid line: jaw; light solid line: tongue body; dashed lines: lips. Time scales are different in each panel.

tongue body (thin solid line) and lips (dashed lines). Phonologically salient effects of movement (e.g. closure) occur at some lag when compared with the activation intervals in the gestural score. In particular, consonantal activation begins well before closure is reached (the point of closure is marked by a solid vertical line), and activation finishes well before consonantal release (the vertical dashed line). Movement traces have been aligned with $t = 0$ at the moment of closure. The motion from a low position or /a/ to a high position for /i/ is evident in both tongue body and jaw traces. The additional movement of the lips after closure is reached represents compression of the soft lip body, as documented, e.g. in Löfqvist and Gracco (1997). It is clear that despite the very different parameter settings employed, and the large difference in overall duration that results, the kinematic form of the /abi/ utterance is essentially stable.

4.1. Context dependency of the relative timing of gestures

In examining the fine detail of gestural sequencing, two kinds of consistency in the face of variation are important. Firstly, we may consider consistency across different segmental sequences, as we change the segmental make up of an utterance. Secondly, consistency in the realization of a single sequence as suprasegmental context changes also provides an important window into the constraints operative in determining fluid, natural movement. With respect to the first, we are in a position to model height variation in vowels, and the specific characteristics of both apical and bilabial consonants. With respect to the latter, we can model prosodic variation along the two dimensions of hypo/hyper-articulation, and speech rate modulation. Examining these within the present modeling framework allows us to consider how both kinds of consistency may arise as a result of generic optimization principles.

Löfqvist and Gracco (1999) investigated the temporal details of sequencing of the tongue and lip movements in asymmetrical VCV sequences with consonants /b/ and /p/ and the vowels /i/, /a/, /u/ uttered by four speakers. They found that that the "onset of the tongue movement from the first to the second vowel almost always occurred before the oral closure" (Löfqvist and Gracco,

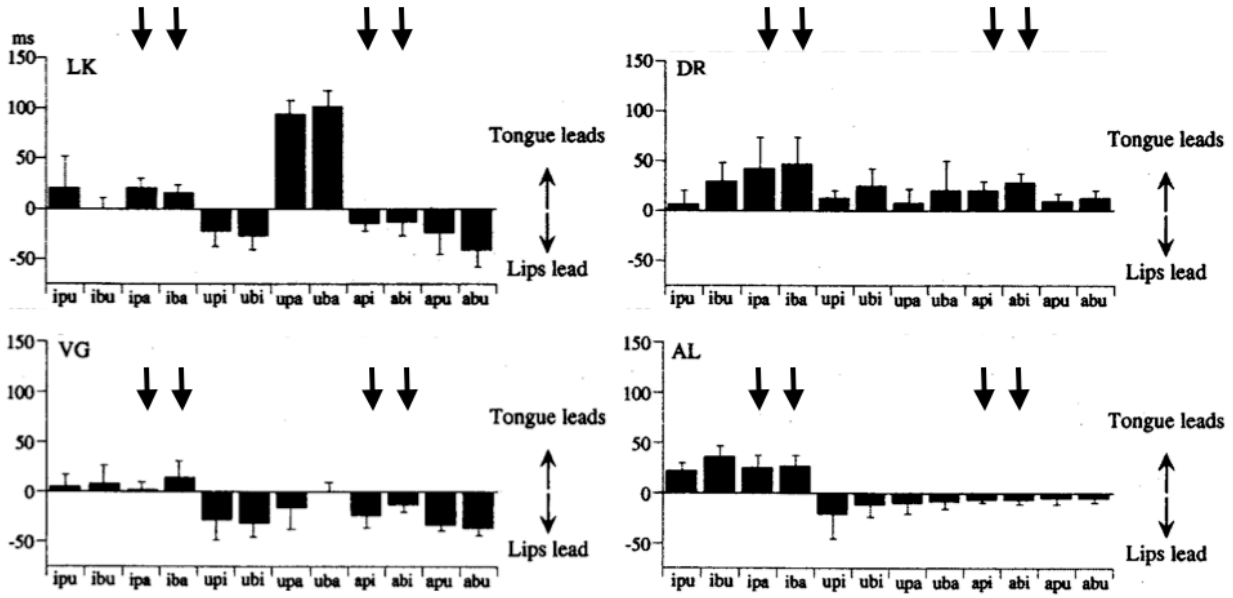

Figure 7: Interval between the onset of the tongue movement and the onset of the lip closing movement for the consonant, from Löfqvist and Gracco (1999). Standard deviations are also plotted. Arrows indicate data discussed in the text.

1999), which suggests a degree of invariance in the sequential order of gestural landmarks. However, the authors also found that the articulatory nature of the first vowel has a reliable influence on the interval between the tongue movement and the oral closure. The tongue movement started relatively earlier before the closure achievement for /iCa/ sequences than for /aCi/ sequences. For sequences containing /u/ the results were less clear and showed a strong speaker dependency, possibly because of a strong influence of lip rounding on the timing of other participating gestures.

Another important issue concerning gestural sequencing is that of the relative timing of the onset of a consonantal bilabial gesture with respect to the intervocalic switch realized by the tongue body. As we have seen, the tongue body movement consistently started before the lip closure was achieved; but so does, necessarily, the lip movement. Two related questions arise. Which of these two movements, indicating the onset of the appropriate gesture, starts earlier? And can we interpret the observed sequencing detail as the hallmark of optimal movement, once we have provided a sufficiently precise definition of "optimal", as within our model?

Fig. 7, from Löfqvist and Gracco (1999) provides some relevant data and suggests an answer to these questions. First, it shows considerable intra- and inter-speaker variability in the relative timing of these two landmarks. More importantly, it also shows variability with respect to the identity of vowels in the given VCV sequence. If we, however, limit our attention to the asymmetric sequences with /a/ and /i/ vowels (indicated by arrows in Fig. 7), whose production is distinguished primarily by the tongue body height, an interesting pattern emerges: for all four subjects the bilabial gesture onset is later than the intervocalic tongue body movement onset in sequences /iba/, /ipa/, while for 3 of 4 speakers the pattern is reversed for sequences /abi/, /api/. Even in the case of

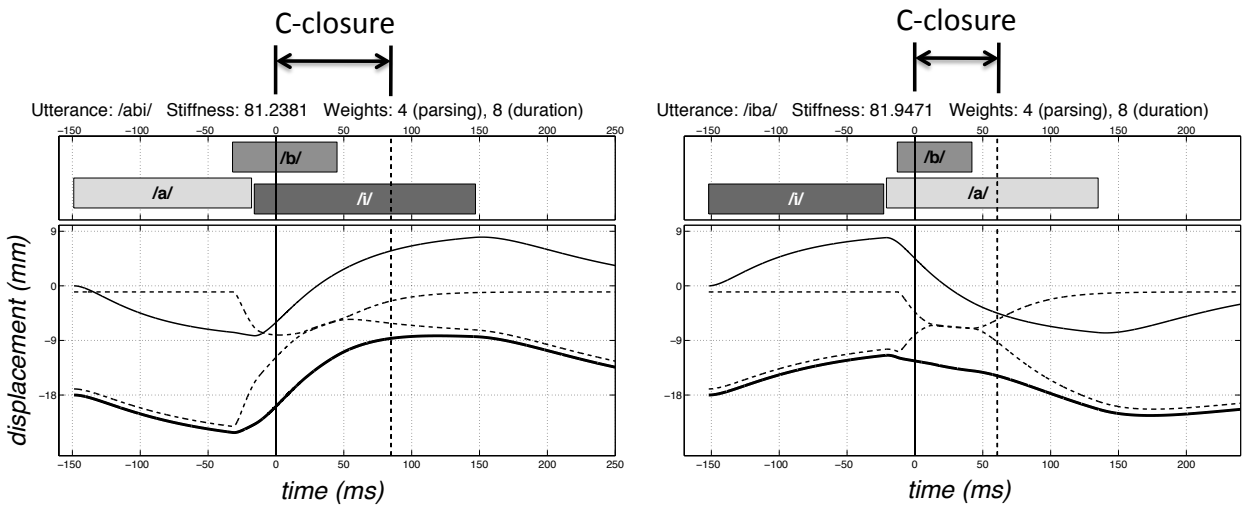

Figure 8: Optimal scores and kinematic traces for /abi/ (left) and /iba/ (right). Mid-range values of 4 for α_P and 8 for α_D were used.

speaker DR, for whom the tongue movement consistently leads the bilabial movement onset, this lead is more pronounced for the sequences starting with a high vowel /i/ than for the sequences /abi/, /api/. Given that the lips are closer together during production of an /i/ vowel than an /a/ vowel, this makes sense. Movement towards closure of the jaw, and hence also the tongue body, can start later for a medial bilabial consonant uttered after /i/ than after /a/, as the distance to be traversed to the point of consonantal closure is smaller.

Model outputs for "moderate" speaking rate and precision requirement ($\alpha_P = 4, \alpha_D = 8$), expressed both as gestural scores, and as associated kinematic traces, for sequences /abi/ and /iba/ are shown in Fig. 8. The optimal constellations discovered by the optimization technique we employ reproduce the qualitative aspects of the observations of Löfqvist and Gracco. Tongue movement onset precedes oral closure (vertical solid line) in each case. Furthermore, consonantal activation occurs slightly before the intervocalic switch for /abi/ but after it for /iba/. Moreover, as we show below, these aspects of the optimal gestural constellation are quite stable for most values of the cost coefficients, although the order of intervocalic switch and consonantal activation onset are reversed for low values of α_P (as in the left hand panel of Fig. 6).

4.2. Search for invariance: Relative phasing

In describing any pattern of coordination, we can express the relative timing of one gesture with respect to the temporal unfolding of another. One well established procedure for doing this is to use the underlying undamped oscillatory cycle of one gesture as a referent for the other, and to express relative timing as phase, ϕ, where $\tan(\phi) = -\dot{x}/x$. This method was introduced in Kelso and Tuller (1985), and has been applied widely since (e.g. Saltzman et al., 2008). The method is motivated by the need to refer to the instantaneous dynamical state of the gesture constellation itself, rather than relating each gesture to an extrinsic time scale. This allows coordinative invariants to be readily expressed, irrespective of changes, e.g. in speaking rate.

Within the Articulatory Phonology framework, the working assumption was made that *fixed*

phasing relations obtained between gestures, and the phase values depended only on the gesture type (V vs C) and its location within the syllable (Browman and Goldstein, 1990b; Saltzman et al., 2008). The phasing relations relevant for our investigations, i.e. the phasing of consonantal gesture with respect to the surrounding vowels, were defined in early accounts of AP (Browman and Goldstein, 1990b) by the following simple rules:

1. A consonant is phased with respect to a preceding vowel so that the target of the consonantal gesture (240°) coincides with a point after the target of the vowel (about 330°)

2. A consonant is phased with respect to a following vowel so that the target of the consonantal gesture (240°) coincides with the onset (0°) of the vowel.

The positing of inflexible and invariant phase relations among gestures was a deliberate simplification, arising out of a desire to abstract away from the contingent biomechanical properties of the vocal tract, and hence to more easily make contact with phonological theory, where such abstraction is the stuff of which theory is built. The authors of the rules stated above readily admit their sketchy and preliminary nature and the somehow arbitrary choice of the precise phase values chosen for the relevant anchor points. Moreover, they admit that they are not particularly confident about the statement (1), as "there is a complex interaction between phasing and stiffness, at least for vowels, about which we still understand very little" (Browman and Goldstein, 1990b). The assumption that the phase relations are invariant with respect to changes in speaking rate and stress patterns was also undermined by subsequent experiments (Nittrouer et al., 1988). We argue below that our modeling paradigm offers important insights into these problems.

As we saw in the previous section, the subsequent detailed articulatory analysis (Löfqvist and Gracco, 1999) revealed a significant dependency of the timing of gestural events on the precise articulatory nature of the speech segments being sequenced. In particular, the presumed *fixed* phase value of the achievement of the consonantal target—a closure—with respect to the underlying abstract cycle of the same consonant does not seem tenable in light of observed, segment-specific variability. The bilabial closure, for example, can be achieved *relatively* earlier when the lips are closer together at the onset of the gestural activation (as in the case of vowel /i/) than when they are further apart (in the case of /a/).

Our simulation results, successfully reproducing Löfqvist and Gracco's (1999) observations, support this intuition. The bilabial closure in the optimal realizations plotted in Fig. 8 is achieved at 224° and 92° of the abstract consonantal cycle for the sequences /abi/ and /iba/, respectively.

Remarkably, however, our simulations do provide some support for Browman and Goldstein's (1990) phasing statements in another important way. In the optimal sequences /abi/, /iba/, /adi/, and /ida/ (for "moderate" setting of the cost weights $\alpha_P = 4, \alpha_D = 16$) generated by our model, the consonantal closure is realized at 284°, 288°, 291° and 284° of the preceding vowel's abstract cycle and at 46°, 43°, 34° and 42° of the abstract cycle of the following vowel, respectively. As we can see, the consonantal targets are realized at stable phase values[1] of the underlying vocalic gestures in a context independent manner, much as postulated by Browman and Goldstein (1990b).

[1]Some of the small variation in the precise phase values is attributable to the limits in precision of the computational implementation of our model. Please note that in the given context 10° of an activation cycle corresponds to less than 5 ms of clock time. Moreover, the exact values of the phases reported here depend on the relatively arbitrary parameter settings of our vocal tract model. However, the stability patterns forming the basis of our argument emerge regardless of the particular setting details.

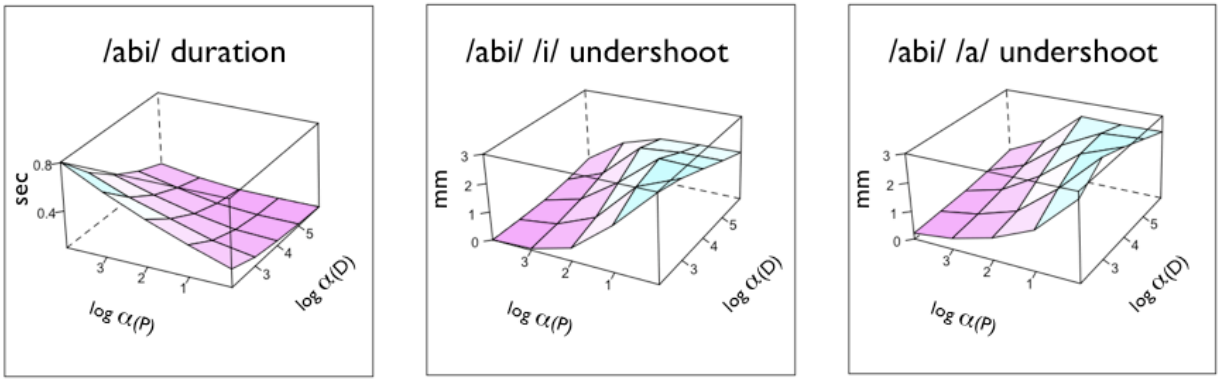

Figure 9: Duration and degree of undershoot for /i/ and /a/ for the sequence /abi/. Cost weights have been log transformed. Plots for other sequences are almost identical.

In the next section, we investigate this invariance further, showing that the form of invariance is different for the relative timing of V1-C and C-V2.

To sum up these results, for fixed values of cost weights, we do not find absolute invariance in phase relations among segments, but we do find a great deal of regularity. Where there are intelligible anatomical grounds for variation as a function of segment identity, we find such variation, which is small in magnitude. But we also find relative stability in the temporal coordination of the consonantal gesture with respect to both preceding and following vowels. This latter finding is emergent, resulting directly from the simple optimization procedure we employ. It is, however, entirely consistent with the intuitions expressed in previous work within AP that sought to codify phasing relations explicitly (Browman and Goldstein, 1990b), or through the use of planning oscillators (Saltzman et al., 2008).

4.3. Invariance and suprasegmental variation

The above stability results were obtained for fixed values of the cost parameters. We can now turn our attention to these phase relations as the cost weights, α_P and α_D, are varied. Variation in these weights is a direct representation of the suprasegmental modulation of clarity of articulation (α_P) and of speech rate (α_D). But before we look into the details of emergent intergestural phasing relations, we must evaluate whether our model is capable of eliciting these intended high level variations.

Fig. 9 shows how overall VCV duration and the degree of undershoot for each of the vowels in the sequence /abi/ vary as the parameters α_P and α_D are varied. Results for the other sequences are nearly identical. It can readily be seen that the optimization procedure produces smooth and readily intelligible variation in both of these resultant variables. Duration varies as a regular function of both parameters (not just α_D, as the parsing cost also affects duration through the definition of realization degree, as in Eqn 6). The degree of hypo-articulation generated is evenly spread across both vowels, and is likewise influenced by both cost terms. Undershoot is not simply inversely related to duration, as the proportion of variance accounted for in one variable by the other (Pearson's r^2) is only 50% for /i/ and 55% for /a/). This accords with Gay (1981), who observed

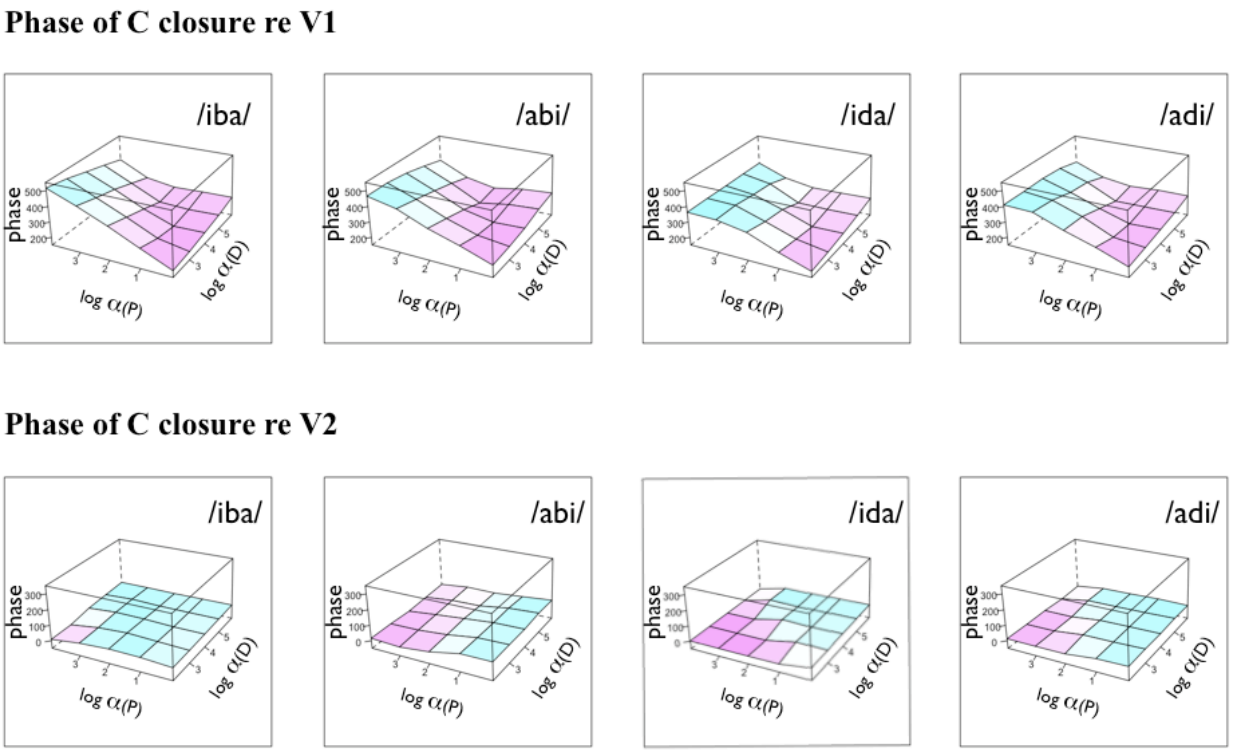

Figure 10: Phase (vertical z-axis) of the medial consonant closure with respect to the (underlying undamped cycle of the) first vowel (top row) and second vowel (bottom row), as the cost component coefficients are varied.

that changes in speaking rate are somewhat independent of the degree of hypo-articulation employed by a speaker in a given utterance. Although Fig. 9 shows a degree of correlation between these two complementary characteristics of an utterance, the model does allow the articulation clarity to vary independently in part of the premium placed on duration. Thus both the undershoot and duration are composite functions of α_P and α_D jointly.

In Browman and Goldstein (1990b), it was suggested that phasing between consonants and vowels was not simple to interpret or predict because of a "complex interplay between phasing and stiffness" (p. 357). Within our optimization framework, the fixed stiffness of each individual component is scaled by the single overall system stiffness value. Crucially, this system stiffness is not controlled, but is emergent, as it is, together with the times of gesture activation onsets and offsets, a variable of the objective function of the optimization procedure. Gestural scores *and* system stiffness are fixed during the evaluation of a gestural score at every single step of the optimization procedure. The interplay between undershoot and duration seen in Fig. 9 is thus mediated by this emergent system stiffness.

We can now turn our attention to the relative timing of the medial consonant and the flanking vowels (Fig. 10), expressed as the phase within the vowel cycle at which C-closure is reached. Looking at the V1-C relation first, for all four sequences, this phase varies systematically and to a large degree as the cost weights change. The relation between the weights and the resulting

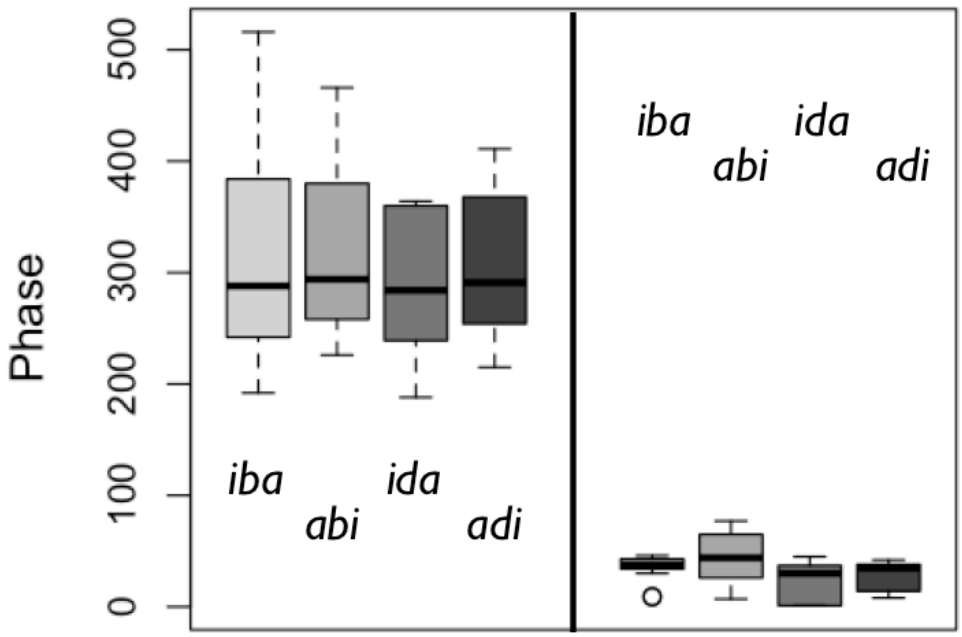

Figure 11: Relative timing of the onset of consonantal closure expressed as a phase of the preceding vowel (left) and of the following vowel (right).

phase is more sensitive to changes in α_P than to α_D, upon which it is only weakly dependent. The consonant closure occurs latest in the vowel cycle for the largest values of α_P, and smallest values of α_D. Within the parametric ranges explored in our simulations, the range of consonant closure phases observed ranges over at least 176 degrees of the underlying vowel cycle (/ida/)[2], to a maximum of 324 degrees (/iba/). Importantly, the effect of cost weight variation on this phase relation is qualitatively the same for all four syllables. Pairwise correlations for V1-C phase range from 0.94 (/ida/ vs /abi/) to 0.99 (/iba/ vs /abi/). The large dependence of the V1-C relation on the cost weight parameters is also readily seen in Fig. 11 (left panel), which shows the distribution of V1-C phases observed.

The C-V2 relation, on the other hand, is essentially invariant across prosodic context. Fig. 10, lower row, shows the phase of the consonant closure with respect to the underlying cycle of the second vowel, while Fig. 11 (right panel) shows the distribution of phases. The variation observed in this phase relation is far less than observed for C-V1, as the cost weights are changed. (Please note: C-V1 is relatively invariant for *fixed* cost weights, C-V2 is relatively invariant *irrespective of* cost weights).

This difference in the stability of the V1-C and V2-C relations is particularly interesting, as this would suggest that unmarked syllabic structure may be yet another emergent arising out of the application of these same simple optimization principles. CV is, as is well known, highly favoured in the world's languages (Ladefoged and Maddieson, 1996), and is also the first syllable form to be found in the emergence of speech from the first stages of babbling (Davis and MacNeilage, 1995).

[2]Phase values are expressed relative to the cycle of the vowel. Values in excess of 360° are possible by extending the cycle into its next period without resetting the phase index.

A similar convergence of articulatory phasing principles with the principles of syllable affiliation found in phonological theories was noted in Browman and Goldstein (1990, p. 357). In contrast to Browman and Goldstein, we do not represent syllable structure anywhere within our model, but instead find evidence for such structure in the pattern of invariant phase relations that emerges.

We have here chosen to examine a wide range of cost weight parameters. It is entirely plausible, indeed likely, that any given individual will employ only a restricted range of variation. It is well known from empirical prosodic research that subjects differ greatly in the amount of speech rate variation they can be induced to exhibit (Cummins, 1999), while a suitable methodology for eliciting substantial variation in the degree of hypo- or hyper-articulation remains to be developed, clear speech studies notwithstanding.

5. Discussion

Fluent, smooth, efficient movement can be regarded as being near optimal with respect to a variety of criteria. The use of second order dynamical systems to model individual gestures generates smooth movement trajectories. One major innovation of the original task dynamic model was to show how this characteristic can be mapped from an abstract, context-independent, task space to the messier set of articulators which may be influenced by several goals simultaneously. Embodied task dynamics, underlying the present model, takes this further by locating the tasks within the same set of physically embodied actuators. Vocal tract elements play a dual role, as end effectors, serving to define the task goals of individual gestures, and as model articulators, subject to multiple simultaneous and competing influences.

But speech presents a far more demanding coordinative challenge than the execution of single, suitably constrained, movements in the service of individual behavioral goals. A complex series of articulatory goals are patterned in overlapping fashion in time. The fluid orchestration of multiple streams of behavioral goals has long represented a challenge to empiricists and theoreticians alike. Articulatory phonology and the conventional task dynamic implementation have already provided a framework within which such fluid movement might arise, but the task of determining an appropriate sequence of gestural activation commands, and the associated problem of modulating system stiffness, have proven to be formidable hurdles to overcome.

A critical opportunity arises with the implementation of an *embodied* task dynamics. Efficiency considerations, and the definition of a composite objective cost function that can be minimized provide a method for coordinating multiple parallel streams of behavioral goals. The cost function we have proposed is no doubt inadequate, but even at this early stage of development, it has allowed us to generate simulated motions of the articulators that appear to have some essential characteristics of natural articulatory data. Timing relations among the gestures are well-behaved with respect to several criteria: the sequencing of articulatory events reproduces many details found in natural data, and it appears to do so in a manner that faithfully represents both the idiosyncratic properties of individual segments, and the more abstract phasing relations that characterize generic properties of VC and CV coordination.

The three components of the proposed objective cost function allow the simple, high-level, exploration of both hypo/hyper-articulation, and speech rate. These are two dimensions of volitional speech control that affect a wide range of details in the resulting speech movement. Without the necessity of postulating very many, individually unmotivated, articulatory rules to account for the articulatory variety that results from natural prosodic modulation, a single principle of optimiza-

tion is applied within this abstract intentional space. A similar argument against the proliferation of rules expressed at the detailed level of individual gestures was presented in Port and Cummins (1992). Browman and Goldstein (1990) managed to express phasing rules that capture much of the flavor of VC and CV coordination (see Section 4.2), but reservations remained about the complex interplay between observed phasing relations and system stiffness. Stiffness, of course, is central to both degree of hypo/hyper-articulation, and to speech rate control, and it interacts with both in subtle ways.

There have been previous attempts to use stiffness as a control parameter, e.g. in Ostry and Munhall (1985). Our approach has been to regard system stiffness as one of the dimensions that are being optimized, along with activation onsets and offsets. In this way, stiffness values appropriate for any given setting of the prosodic parameters fall out of the model quite naturally. This leads us to suggest that the two rules proposed in Browman and Goldstein (1990b, Section 4.2 herein) might serve as approximate descriptions of the phasing relations that obtain between vowels and consonants. The phasing of a vowel with a following consonant then appears at a relatively fixed value *for specific fixed values of the parameters of suprasegmental variation* (Fig. 10, top row), while the phasing of a consonant with a following vowel appears to be approximately invariant, regardless of prosodic modulation for most of the ranges explored herein (Fig. 10, bottom row).

Within our approach, two salient characteristics of movement are inextricably linked. On the one hand, our optimization procedure serves to generate fully specified movement trajectories, rich in kinematic detail that can be interpreted, e.g. as exhibiting varying degrees of under- or overshoot. On the other, the same procedure provides an explicit account of the relative timing patterns among individual gestures. In its ability to make specific kinematic predictions, therefore, our model exhibits similarities to the DIVA model (Guenther, 1995; Guenther et al., 1998), which also seeks to provide an account of such detail in movement as, for example, the relation between the magnitude of undershoot and associated rate changes. This is approached within DIVA by varying the size of the articulatory targets as the rate control signal varies. DIVA, however, does not attempt to provide any precision in accounting for intergestural sequencing relations, as demanded here.

The concepts employed in our model obviously bear a strong similarity to the Task Dynamic implementation of Articulatory Phonology, which was, in many respects, its starting point. Two major points of divergence can be seen in the Embodied Task Dynamic model. Firstly, the entire modeling framework has changed from providing an account of *online* control and constraint in movement to providing an account of why one form of movement is to be preferred over another, and justifying that differentiation using explicit criteria of optimality. Uncovering the optimality landscapes underlying gestural sequencing is a computationally expensive procedure, and is not suited to the development of an online control algorithm.

The second principal point of divergence serves to justify our model's name as *Embodied* Task Dynamics. In our model, dynamical systems are inseparable from the articulatory means with which the gestures are realized. We thus find an interplay between dynamical landmarks, such as phasing relations expressed as fixed proportions of a limit cycle, and articulatory constraints, such as the effect of the collision of an articulator with a fixed vocal tract boundary. To provide a concrete example, the stable phasing relation between the intervocalic consonant and the subsequent vowel that emerged from our simulations serves to link a *dynamic* attribute of the vocalic gestural cycle—a fixed phase value—with an *articulatory* event—the associated consonantal closure. This interplay between dynamical and physical events is facilitated by the very nature of the embodied

task dynamics deployed in our model.

Two obvious extensions of this work are now clear. Firstly, empirical data on the effect of two dimensions of prosodic variation on coordinative timing and gestural amplitude need to be obtained and compared with the performance of the model. Articulatory or acoustic recordings do not provide a direct access to the precise values of inter-gestural phasing relations, but can yield measurable data on macroscopic patterns of gestural sequencing (the order in which the gestures are triggered). The relationships between various kinematic characteristics of recorded articulatory material—such as the amount of undershoot, durational properties, velocity profiles, etc.—can be compared to the predictions generated by our model. We have explored a wide range of variation in both hypo/hyper-articulation and in speech rate modulation. Any given subject may be reasonably expected to display only a restricted range of variation in either of these. Secondly, and subsequently, the simple and highly restricted vocal tract geometry we have employed needs to be extended. The addition of a velum tract variable poses no particular problem, as the velum is anatomically relatively independent of the other articulators. A greater and more important task lies in moving from a one dimensional vowel space to a two dimensional space. If the model continues to behave well with that elaboration, it can reasonably be further evaluated within a full articulatory synthesis system.

References

Abbs, J. H. and Gracco, V. L. (1983). Sensorimotor actions in the control of multimovement speech gestures. *Trends in Neuroscience*, 6:393–395.

Anderson, F. C. and Pandy, M. G. (2001). Dynamic optimization of human walking. *ASME Journal of Biomechanical Engineering*, 123:381–390.

Bernstein, N. (1967). *The Coordination and Regulation of Movements*. Pergamon Press, London.

Biewener, A. and Taylor, C. (1986). Bone strain: a determinant of gait and speed? *The Journal of Experimental Biology*, 123:383–400.

Browman, C. P. and Goldstein, L. (1990a). Articulatory gestures as phonological units. *Phonology*, 6:201–251.

Browman, C. P. and Goldstein, L. (1990b). Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J. and Beckman, M. E., editors, *Between the Grammar and Physics of Speech: Papers in Laboratory Phonology I*, pages 341–376. CUP, Cambridge.

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

Byrd, D. (1996). A phase window framework for articulatory timing. *Phonology*, 13:139–169.

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

Davis, B. and MacNeilage, P. (1995). The articulatory basis of babbling. *Journal of Speech, Language and Hearing Research*, 38(6):1199–1211.

Fitts, P. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology: General*, 47:381–391.

Fowler, C. A., Rubin, P., Remez, R., and Turvey, M. (1981). Implications for speech production of a general theory of action. In Butterworth, B., editor, *Language Production*, pages 373–420. Academic Press, San Diego, CA.

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica*, 38:148–158.

Gray, G. (1942). Phonemic microtomy: The minimum duration of perceptible speech sounds. *Communication Monographs*, 9(1):75–90.

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

Guenther, F. H., Hampson, M., and Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105:611–633.

Harris, K. S. (1987). Action theory as a description of the speech process. In Peters, H. F. M. and Hulstijn, W., editors, *Speech Motor Dynamics in Stuttering*, chapter 2, pages 25–39. Springer, New York.

Hogan, N. and Flash, T. (1987). Moving gracefully: quantitative theories of motor coordination. *Trends in Neurosciences*, 10(4):170–174.

Hoyt, D. and Taylor, C. (1981). Gait and the energetics of locomotion in horses. *Nature*, 292(5820):239–240.

Jordan, M., Flash, T., and Arnon, Y. (1994). A model of the learning of arm trajectories from spatial deviations. *Journal of Cognitive Neuroscience*, 6(4):359–376.

Kay, B., Saltzman, E., and Kelso, J. A. S. (1991). Steady-state and perturbed rhythmical movements: Dynamical modeling using a variety of analytical tools. *Journal of Experimental Psychology: Human Perception and Performance*, 17:183–197.

Keller, E. (1987). The variation of absolute and relative measures of speech activity. *Journal of Phonetics*, 15:335–347.

Kelso, J. A. S. and Tuller, B. (1987). Intrinsic time in speech production: theory, methodology, and preliminary observations. In Keller, E. and Gopnik, M., editors, *Motor and Sensory Processes of Language*, pages 203–222. Lawrence Erlbaum Associates Inc, Hillsdale, NJ.

Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the World's Languages*. Wiley-Blackwell.

Lashley, K. S. (1951). The problem of serial order in behavior. In Jefress, L. A., editor, *Cerebral Mechanisms in Behavior*, pages 112–136. John Wiley and Sons, New York, NY.

Lindblom, B. (1983). Economy of speech gestures. In MacNielage, P., editor, *The Production of Speech*, pages 217–245. Springer Verlag, New York.

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.

Lindblom, B. (1999). Emergent phonology. In Chang, S., Liaw, L., and Ruppenhofer, J., editors, *Proc. 25th Annual Meeting of the Berkeley Linguistics Society*, pages 195–209, Berkeley, CA. University of California.

Lindblom, B., Guion, S., Hura, S., Moon, S.-J., and Willerman, R. (1995). Is sound change adaptive? *Rivista Di Linguistica*, 7(1):5–37.

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

Nakano, E., Imamizu, H., Osu, R., Uno, Y., Gomi, H., Yoshioka, T., and Kawato, M. (1999). Quantitative examinations of internal representations for arm trajectory planning: Minimum commanded torque change model. *Journal of Neurophysiology*, 81(5):2140–2155.

Nam, H. and Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2253–2256, Barcelona, ES.

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

Ohala, J. J. (1989). Discussion of Björn Lindblom's "Phonetic Invariance and the Adaptive Nature of Speech". In Bouma, H. and Elsedoorn, B., editors, *Working Models of Human Perception*, pages 175–183. Academic Press, London, UK.

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

Ostry, D. J. and Munhall, K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, 77(2):640–648.

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

Port, R. and Cummins, F. (1992). The English voicing contrast as velocity perturbation. In Ohala, J., Nearey, T., Derwing, B., Hodge, M., and Wiebe, G., editors, *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 1311–1314. University of Alberta.

Rubin, P. E., Baer, T., and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70:321–328.

Saltzman, E. and Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review*, 94:84–106.

Saltzman, E. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382.

Saltzman, E., Nam, H., Krivokapic, J., and Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In Barbosa, P. A., Madureira, S., and Reis, C., editors, *Proceedings of the 4th International Conference on Speech Prosody.*, Campinas, BR.

Schwartz, J.-L., Boe, L.-J., Vallee, N., and Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25:255–286.

Simko, J. (2009). *The Embodied Modelling of Gestural Sequencing in Speech*. PhD thesis, UCD School of Computer Science and Informatics, University College Dublin. Also released as Technical Report UCD-CSI-2009-07 available from http://www.csi.ucd.ie/biblio.

Simko, J. and Cummins, F. (2010). Embodied task dynamics. *Psychological Review*. In Press.

Srinivasan, M. and Ruina, A. (2006). Computer optimization of a minimal biped model discovers walking and running. *Nature*, 439(7072):72–75.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, 7:907–915.

Todorov, E. and Jordan, M. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235.

Uno, Y., Kawato, M., and Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement. *Biological Cybernetics*, 61(2):89–101.

van Son, R. J. J. H. and Pols, L. C. W. (2002). Evidence for efficiency in vowel production. In *Proceedings of ICSLP*, Denver, USA.

van Son, R. J. J. H. and Pols, L. C. W. (2003). How efficient is speech? In *Proceedings of the Institute of Phonetic Sciences*, volume 25, pages 171–184, University of Amsterdam.

Wolpert, D., Ghahramani, Z., and Flanagan, J. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, 5(11):487–494.