Embodied Task Dynamics

Juraj Simko and Fred Cummins*
UCD School of Computer Science and Informatics
University College Dublin
{*juraj.simko,fred.cummins*}*@ucd.ie*

September 6, 2010

Preprint of article to appear in *Psychological Review*, 2010. There may be minor textual differences between this informal preprint and the publication of record.

* To whom correspondence should be addressed.

Abstract

Movement science faces the challenge of reconciling parallel sequences of discrete behavioral goals with observed fluid, context-sensitive motion. This challenge arises with a vengeance in the speech domain, where gestural primitives play the role of discrete goals. The task dynamic framework has proved effective in modeling the manner in which the gestural primitives of articulatory phonology can result in smooth, biologically plausible, movement of model articulators. We present a variant of the task dynamic model with one significant innovation: tasks are not abstract and context-free, but are embodied and tied to specific effectors. An advantage of this approach is that it allows the definition of a parametric cost function which can be optimized. Optimization generates gestural scores in which the relative timing of gestures is fully specified. We demonstrate that movements generated in an optimal manner are phonetically plausible. Highly nuanced movement trajectories are emergent based on relatively simple optimality criteria. This addresses a long standing need within this theoretical framework, and provides a rich modeling foundation for subsequent work.

Keywords: Task Dynamics, Articulatory Phonology, Embodiment, Coordination, Sequencing

Introduction

Intentional movement can be expressed in terms of high-level discrete behavioral goals, such as reaching for a cup, ascending stairs, or reciting a poem. Any such description may admit of smaller parts, such as speaking each word of the poem, each syllable of each word, or each individual articulatory gesture within each syllable. Movement trajectories that result are smooth, continuous, and fluid, accommodating contextual influences of overlapping goals and environmental variation. This sets the stage for a very basic problem in understanding action: how are discrete behavioral goals realized within a very high-dimensional bio-mechanical system, such that the resulting movements unfold smoothly, in a context-sensitive manner.

Several models, including the task dynamic model, address this question by proposing some form of control algorithm (Saltzman and Kelso, 1987; Saltzman and Munhall, 1989; Guenther, 1995). We take an alternative stance, and seek to explain the observed temporal patterning of movement as the result of the satisfaction of multiple criteria of optimality. We employ computational techniques that differentiate between more and less efficient forms of movement, but we do not propose these as on-line control procedures. Rather, we arrive at a precise formulation of the relationship between high-level intentional parameters, such as speech rate or the degree of hyper- or hypo-articulation, and the corresponding form of movement that is deemed optimal, given those settings.

The theory of Articulatory Phonology occupies a unique position within linguistics as it seeks to provide a unified account of the patterning of movement and sound in language that can do duty in the fields of both phonology and phonetics. Couching the primitives of a phonological theory in terms of movement goes a long way towards de-mystifying the relationship between the apparently rule-based world of discrete sound patterns familiar from phonology and the smooth, continuous, and physically instantiated world of phonetics (Fowler et al., 1980). A good introduction to Articulatory Phonology can be found in Browman and Goldstein (1992) and also in Browman and Goldstein (1995), which is accompanied, in the same volume, by a description of the task dynamic implementation that has become inseparably associated with Articulatory Phonology (Saltzman, 1995).

The task dynamic framework was originally formulated to help in the parsimonious and explicit modeling of smooth skilled movement (Saltzman and Kelso, 1987), and later applied to speech (Saltzman and Munhall, 1989). An informal overview is provided in Hawkins (1992). Task dynamics seeks to provide a description of movement that acknowledges both context-free discrete behavioral units, such as reaching for a cup, or uttering a vowel, and continuous, context-dependent co-production as found in kinematic traces of the limbs or articulators. Tasks are specific goals that are expressed at the level of meaningful discrete units. In speech, these correspond, essentially, to the gestural primitives of Articulatory Phonology. An example of a task might be to form and hold a bilabial closure, or to move the tongue towards a position appropriate for producing an /a/ vowel.

Each task is modeled as a context-free, second-order mass-spring dynamical system with critical damping. This somewhat dense description means that at the task level, each task is completely independent of the others. It is expressed as a simple abstract point mass that starts at some remove from its goal, and then moves in a smooth fashion directly to the goal state, in accordance with the dynamic expressed in the differential equations for the task. Critical damping ensures that the goal state is neither under- nor over-shot in the approach. In modeling speech, the position of the abstract mass relative to the intended target constitutes a *tract variable*. Context independent tract variable trajectories then map into a set of articulators. Several tract variables can vie for

control of a single articulator and a single task can seek to influence multiple articulators. Much of the mechanics of the task dynamic model is concerned with ensuring that these mappings from tract space to articulator space are conducted in a way that ensures that smooth, realistic, context-sensitive movement trajectories result.

Within Articulatory Phonology, an utterance is specified using a gestural score (Figure 1). The solid blocks of the score specify the time intervals in which specific tasks are active. The continuous traces show the resultant motion of the tract variables. These, in turn, are mapped directly into the space of model articulators. In Figure 1, each row corresponds to one tract variable. Some articulators are associated with more than one tract variable. For example, the jaw participates in motion associated with the lips, tongue body and tongue tip. Likewise, some tract variables are associated with several articulators. Here, the lip aperture tract variable is mapped to the motions of the upper and lower lips and the jaw. The resulting movement traces can be used to parameterize an articulatory synthesizer to generate sound output (Rubin et al., 1981), though the generation of appropriate movements is the proximal goal of the task dynamic implementation. The Articulatory Phonology framework has provided many insightful analyses of articulatory phenomena, including consonant and vowel coarticulation and separation, apparent consonant deletion and assimilation in casual speech, segment insertions, etc (Browman and Goldstein, 1990).

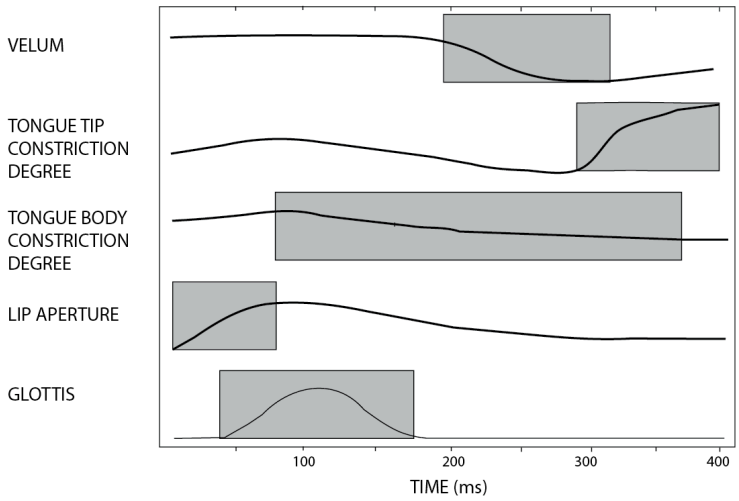

Figure 1: Partial gestural score for the utterance /pan/ and resulting tract variable trajectories. Adapted from Browman and Goldstein (1995).

One outstanding problem has been the specification of the temporal details of the activation intervals in the gestural score. Clearly, very many features of the resulting movement will depend critically on the exact timing of the gestures relative to each other. The details of movement evolution in time depend critically on the periods during which gestures are active, and also on the stiffnesses of the individual model components. Stiffness, here, is a standard part of a mass-spring model which is used to implement the gestures. Stiffer components move more rapidly, but also at an increased metabolic cost. In early work, gestural scores were specified by *fiat*. Attempts were later made to derive relative timing through the use of neural networks, or by specification of 'phase windows' within which timing could vary (Byrd, 1996). More recently, sequencing has

been addressed using coupled oscillator models, with some success (Saltzman and Byrd, 2000). But the principled determination of appropriate sequencing and relative timing of gestures within the model remains a major research goal, and is of both practical and theoretical significance.

A further criticism of the task dynamic model in particular has been raised by Hawkins:

> [M]ass is always abstract in the task-dynamic system, but for some articulations the real mass of an articulator is crucial. If it turns out that physical mass does have to be used to account for sounds like trills and apical vs. laminal stops, then it would be reasonable to reevaluate the relationship between abstract and physical mass. (Hawkins, 1992, p. 20)

In this paper, we present a modification of the task dynamic model which addresses both of these concerns. Firstly, we alter the specification of the task space, relocating it in the physical space of the articulators. Tasks are now seen as dynamical systems with a dynamic defined over real masses, such as the jaw, the tongue, etc. This innovation motivates our use of the term Embodied Task Dynamics. The modification is non-trivial, and has as a consequence an inevitable coupling among tasks in task space—something that task dynamics has hitherto shunned. A significant advantage of having physically embodied tasks is then presented. Because the tasks are embodied, it is possible to define physical costs for movement. A parametric cost function is presented that consists of three weighted components. The components implement straightforward constraints related to efficiency and effectiveness in communication. By optimizing relative timing and system stiffness with respect to this cost function, it is possible to derive optimal relative timing among the various elements of the gestural score. This in turn gives rise to smooth, phonetically plausible, movement trajectories.

Because the introduction of embodiment into the task space of the task dynamic model poses some technical challenges, this paper will limit itself to establishing the embodied task dynamic architecture and model, and will limit its treatment of the role of optimization to just that which is necessary to appreciate the consequences of this innovation. A subsequent article will attend in much greater detail to the optimization process, its sub parts, and its consequences (Simko and Cummins, 2010). A summary overview of the goals of the model, and the purpose of optimization are provided in Simko and Cummins (2009).

Before going any further, a brief discussion of the distinction between a *dynamic* description of movement, and a *kinematic* description of the same movement, is in order. Dynamics describes motion as it arises from forces. The Newtonian description of force is the product of mass and acceleration. Masses constrain movement, and likewise, inertial properties ensure a smoothness to movement that is characteristic of the movement of real objects, both animate and inanimate. A kinematic description does not make reference to physical forces, and is not subject to any such constraint. Cartoon animators are free to vary the kinematics of the movements of their characters at will, thereby apparently violating the laws of physics. When we here discuss the origin of movement within our model and other models, we will attempt to be clear about whether that movement arises from equations essentially involving masses and Newtonian constraints on the movement of those masses, or not. As will be come clear, masses and inertial constraints play rather different roles in the original task dynamic model and in our embodied development thereof.

Task Dynamics with Abstract Tasks

We first provide an overview of the task dynamic model as introduced in Saltzman and Munhall (1989). For full details, the reader should consult Saltzman and Kelso (1987), Saltzman and Munhall (1989) and Saltzman (1991). Within this approach, a gesture (the atomistic unit of Articulatory Phonology) is a behavioral task. Examples include the bilabial constriction required for a /b/ sound, a tongue configuration suitable for an /a/, or the lowering of the velum. The first two of these involve composite goals of making a constriction of a specific *degree* in a specific *place*. Each primitive goal has an associated *tract variable*: so to produce an /a/ there are two tract variables: Tongue Dorsum Constriction Degree and Tongue Dorsum Constriction Location. For the velum lowering, there is a single associated tract variable (Velic Aperture).

When a given gesture is active (as specified within the gestural score; see Figure 1), each associated tract variable exhibits change over time, as specified by its own differential equation (a critically damped mass-spring system). Starting from some position, the tract variable moves smoothly towards its target, z_0, just as the restoring force exerted by a stretched spring will move an attached mass back to its equilibrium position.

During a gestural activation interval, the behavior of the tract variables, collectively represented by the vector $\mathbf{z} = (z_1, \ldots, z_n)^T$, is governed by the system of differential equations

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

For the sake of concreteness, we set the number of tract variables $n = 8$. The diagonal 8×8 matrices $\mathbf{M} = \mathrm{diag}(m_1, \ldots, m_8)$, $\mathbf{K} = \mathrm{diag}(k_1, \ldots, k_8)$ and $\mathbf{B} = \mathrm{diag}(b_1, \ldots, b_8)$ and the vector $\mathbf{z}_0 = (z_{01}, \ldots, z_{08})^T$ contain the mass, stiffness, damping, and equilibrium point parameters of the mass-spring tract variable system.

As conventionally implemented, task dynamics presumes the mass parameter of the dynamics of each abstract tract variable to be arbitrary, i.e., not related to any real masses acted upon by the speech production system. Therefore, $\mathbf{M}$ is simply set to be the 8×8 diagonal identity matrix, with each $m_i = 1$. So, the tract variable dynamics is, in fact, defined by an even simpler system:

$$\ddot{\mathbf{z}} = -\mathbf{K}(\mathbf{z} - \mathbf{z_0}) - \mathbf{B}\dot{\mathbf{z}} \tag{1a}$$

where the redundant mass parameter of the system dynamics does not have to be represented at all.

The spring-like dynamics is modeled as critically damped to avoid oscillations. The values of the damping coefficients are thus analytically related to the stiffness and the mass parameters: $b_i = 2\sqrt{m_i k_i}$. The matrix $\mathbf{B}$ can be expressed as a function of the matrices $\mathbf{M}$ and $\mathbf{K}$.

Therefore, during the gestural activation interval, the behavior of each tract variable z_i is in fact determined by two parameters only: the corresponding equilibrium point value z_{0i} representing the gestural target (constriction degree or location) and the corresponding value of the stiffness matrix ($\mathbf{K}$) diagonal determining the velocity with which the system proceeds to achieve the given task. These parameters are *assigned* to a given gesture and do not correspond to any physiological properties of the vocal tract.

The matrices $\mathbf{M}$, $\mathbf{K}$, and $\mathbf{B}$ being diagonal, the constituent equations of the dynamical system given in Equation 1 are uncoupled. The tract variables are thus assumed to represent independent modes of articulatory behavior that do not interact dynamically.

Kinematics of model articulators

The task dynamic implementation of Articulatory Phonology provides a rigorous definition of the system's behavior at the task, or tract variable, level. In addition, it provides a mathematical means for translating between various, related, levels of description. In particular, Saltzman and Munhall (1989) provide a mechanism for translating from the dynamics of tract variables to the kinematics (movements) of the model articulators. The tract variables, whose trajectories are illustrated in Figure 1, are abstract and context-free. The mapping to model articulators is potentially many-to-many, as illustrated in Table 1.

	LH y_1	JA y_2	ULV y_3	LLV y_4	TBR y_5	TBA y_6	TTR y_7	TTA y_8
LP z_1	•							
LA z_2		•	•	•				
TTCL z_3		•			•	•	•	•
TTCD z_4		•			•	•	•	•
TBCL z_5		•			•	•		
TBCD z_6		•			•	•		

Table 1: Matrix representing the relationship between some tract variables (zs, rows) and articulatory variables (ys, columns), adapted from Saltzman and Munhall (1989). The dots in a given tract-variable row indicate that the corresponding articulators contribute to the tract-variable's motion. Articulators (columns): LH: Horizontal lip movement; JA: Jaw angle; ULV/LLV: Upper (lower) lip vertical displacement; TBR/TBA: Tongue body radial and angular position; TTR/TTA: Tongue tip radial and angular displacement. Tract variables (rows): LP: Lip protrusion; LA: Lip aperture; TTCL/TTCD: Tongue tip constriction location and degree; TBCL/TBCD: Tongue body constriction location and degree.

Table 1 provides an outline of the complex relationship between (a subset of) tract variables and model articulators. It shows that the motion of almost every tract variable can be influenced by the movement of several model articulators. The tongue tip constriction degree tract variable ($TTCD$), for example, is linked to the model articulator variables associated with the tongue tip (tongue tip radial and angular positions TTR and TTA). At the same time it also depends, albeit indirectly, on the behavior of the tongue body (articulatory variables TBR and TBA) and the jaw (JA).

The many-to-one mapping of the values of the articulatory variables (collectively expressed by the vector $\mathbf{y}$ of the values of all model articulator variables[1]) to the task variable space is thus redundant. The speech production system makes use of this redundancy when compensating for perturbations or external restrictions imposed on individual speech articulators. On the other hand, this redundancy poses a problem for the modeler: the reverse transformation of tract variable values to articulator positions is under-determined—there is a continuum of articulatory constellations all yielding the same values of a given tract variable.

To solve this problem, Saltzman and Munhall (1989) proposed a kinematic projection of tract

[1]We consistently use z to denote tract variables, and y for the model articulators. The latter variables are labelled θ in the original TD model.

variable accelerations to model articulator accelerations as follows: First the redundant direct (or forward) kinematic mappings from model-articulators to tract-variables are made explicit (Equations 2–4 below). Since the inverse mapping from tract variable accelerations to model articulator accelerations is underdetermined, the Jacobian pseudo-inverse is used to provide an optimal least-squared pattern of model articulator accelerations

As established above (and illustrated by Table 1), the model tract variables can be expressed as functions of the corresponding model articulators. In vector form, this mapping can be expressed as:

$$\mathbf{z} = \mathbf{z}(\mathbf{y}). \tag{2}$$

where the exact form of this function is determined by the particular vocal tract geometry that is employed.

The following direct kinematic relationships then hold for the first and second time derivatives of the mapping $\mathbf{z}$:

$$\dot{\mathbf{z}} = \mathbf{J}(\mathbf{y})\dot{\mathbf{y}} \tag{3}$$

$$\ddot{\mathbf{z}} = \mathbf{J}(\mathbf{y})\ddot{\mathbf{y}} + \dot{\mathbf{J}}(\mathbf{y}, \dot{\mathbf{y}})\dot{\mathbf{y}}, \tag{4}$$

where $\mathbf{J}(\mathbf{y})$ is the Jacobian transformation matrix of the mapping $\mathbf{z} = \mathbf{z}(\mathbf{y})$ whose elements J_{ij} are partial derivatives $\partial z_i / \partial y_j$ evaluated at the current $\mathbf{y}$.

Using the relationships in Equations 2–4, the tract variable dynamics (Equation 1a) can be recast to the model articulator variable space, and the articulatory acceleration vector $\ddot{\mathbf{y}}_A$ representing the active driving influences on the model articulators can be expressed as

$$\ddot{\mathbf{y}}_A = \mathbf{J}^*(-\mathbf{K}\triangle\mathbf{z} - \mathbf{B}\mathbf{J}\dot{\mathbf{y}}) - \mathbf{J}^*\dot{\mathbf{J}}\dot{\mathbf{y}}, \tag{5}$$

where $\triangle\mathbf{z} = \mathbf{z}(\mathbf{y}) - \mathbf{z}_0$ is the distance vector of active tract variables from the given set of targets $\mathbf{z}_0$, and $\mathbf{J}^* = \mathbf{W}^{-1}\mathbf{J}^T(\mathbf{J}\mathbf{W}^{-1}\mathbf{J}^T)^{-1}$ is a weighted pseudo-inverse of the Jacobian transformation $\mathbf{J}$; $\mathbf{W}$ being an appropriate diagonal weight matrix. The derivation and use of a pseudo-inverse in this manner is described, e.g. in Klein and Huang (1983). The matrix $\mathbf{W}^{-1}$ apportions relative weights to the articulators, thereby establishing a pattern of relative "receptivities" among them to the driving influences generated by the tract variables. Importantly, Equation 5 introduces a task dependent *coupling* among the dynamics of the individual articulators. While the changes of individual tract variables over time are mutually independent, the articulators exert reciprocal and continuous influence upon each other.

It is important to note that the model articulators are defined in strictly kinematic terms; they have lengths but no masses. Thus, the coordinate transformation expressed by Equations 2–4 is a kinematic one (Saltzman, 1991), transforming the system state description from an articulatory frame of reference ($\mathbf{y}$) to the task space ($\mathbf{z}$). The articulatory dynamics defined by Equation 5 does not make the physical properties (masses, stiffness, damping) of the articulators explicit. For redundant systems, this shortcoming can, to an extent, be mitigated during the reverse mapping from task space ($\mathbf{z}$) to articulator space ($\mathbf{y}$) by assigning appropriate values to the diagonal of the weight matrix W used for the Jacobian pseudo-inverse computation. Indeed, when the values on the diagonal correspond to the masses acted upon by the articulators, the dynamic behavior of the articulators becomes scaled with respect to these masses: the heavier the articulator, the less receptive it is to the task-evoked action of the vocal tract.

As mentioned earlier, the main focus of this paper is to present a platform for capturing the influence of the physiological properties of individual articulators on resulting movement patterns and timing of speech gestures. In order to do that, we need to clearly separate the dynamic effects imposed upon the system by the active tasks from the consequences of the embodied nature of the vocal tract. In other words, we need to embody the task oriented action, so that the kinematics of each articulator is straightforwardly attributable to the action of forces upon its mass.

The manner in which the relative influence of the masses upon articulatory kinematics is distributed using the pseudo-inverse weight technique described above does not fully satisfy this requirement. The forces applied to the articulators are pre-scaled in a task dependent fashion to reflect the mass distribution. The dynamics of every articulator reflects the relation of its own mass with respect to the masses of all currently active articulators engaged in realizing the simultaneously active tasks. The influence of the task and that of the physiological properties of the given articulator are inseparable. Moreover, as we show in section Embodied Task Dynamics: The Dynamics, this possibility of scaling the articulator behavior with respect to masses completely disappears when the system is further constrained by additional tasks. In traditional TD, the articulator behavior is constrained by the theoretical assumption of the uncoupled, independent nature of active gestural dynamics, which is in accord with the assumptions of Articulatory Phonology.

In the next section, we shall adapt Equation 5 so that the resulting dynamics driving our model more transparently reflects the physiological properties of model articulators in a task independent fashion. The basic idea of our approach is to refine the parameterization so that the dynamical parameters of the model articulator system remain interpretable in the appropriate domains of description. The stiffness and the equilibrium position parameters, imposing the task oriented behavior, remain defined within the tract variable (end-effector) coordinate system. At the same time, the mass parameters representing embodied physical properties of model articulators, are specified at the level of model articulators. The influence of such parameterization on the overall system's dynamics is thus bi-directional—the task dynamics induces a coupling among the articulators, and the embodied articulatory dynamics induces a coupling among the tract variables, reflecting the physical properties of the articulators engaged in given tasks. Computationally, this means that the matrix expressing stiffnesses ($\mathbf{K}_z$) is diagonal in the equations determining the dynamics of tract variables, while the matrix describing articulator masses ($\mathbf{M}_y$) is diagonal in the equations determining articulator movement.

Neutral Attractor, Gating, and Blending

In a task dynamic implementation, each articulator is influenced by zero, one or several tract variables. When no tract variable is influencing its movement, a separate dynamical regime is implemented, called the neutral attractor. This ensures that the articulator relaxes back to a resting position. A gating mechanism ensures that either the neutral attractor or a non-empty set of tract variables influences the articulator, but not both. Full details are in Saltzman and Munhall (1989).

When multiple active gestures compete to influence the same tract-variable, their respective influences are blended, as detailed in Saltzman and Munhall (1989). This involves specifying a weighting scheme for blending the dynamic parameters associated with these gestures. The parameters for which compromise values need to be calculated are stiffness ($\mathbf{K}$), damping ($\mathbf{B}$), equilibrium positions ($\mathbf{z_0}$) and the weight matrix ($\mathbf{W}$) used in Equation 5.

Some changes to the standard implementation of the neutral attractor, gating and blending will

be necessary in an embodied implementation.

Embodied Task Dynamics: The Basic Model

We work with a greatly simplified vocal tract model which comprises a highly restricted set of model articulators (Figure 2). Articulator movement is restricted to a single vertical dimension, allowing contrasting tongue body positions for the vowels /i/ and /a/. The tongue tip is attached to the tongue body, which in turn is yoked to the jaw. The lower lip is also attached to the jaw, while the upper lip has a fixed point of support[2]. These individual components, providing linkages between functionally relevant parts of the vocal tract, are called *pure articulators*. The tongue tip pure articulator thus represents the position of the tongue tip with respect to the tongue body. At present, no glottal or velar modeling is done.

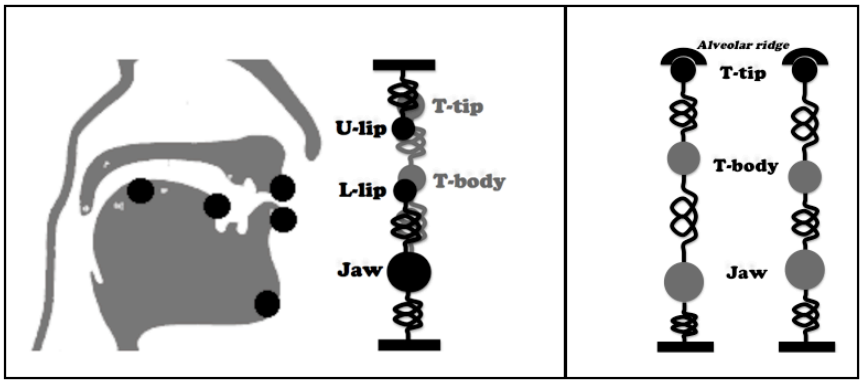

Figure 2: Left: 2-D mid-saggital section, and a corresponding 1-D projection, in which all movements are in the vertical dimension only. Right: Two functionally equivalent configurations of end effectors and pure articulators realising alveolar closure with the tongue tip end-effector.

Functionally relevant parts of active articulators are called the *end effectors* of the goal oriented action. Figure 2 (right panel) illustrates two functionally equivalent articulator configurations. In each, the end effector, here the tongue tip, is fully extended to the alveolar ridge. In this instance, end effector position is a function of three articulators, jaw, tongue body and tongue tip, which here occupy different static configurations in reaching the same target state. The articulators are shown as balls, with size proportional to their respective masses. In contrast to the pure articulator descriptor, the tongue tip end effector position is the absolute location of the tongue tip within the vocal tract boundaries. It is convenient to think of a pure articulator as a mass-spring system, whose position corresponds to the location of its mass relative to its spring's proximal attachment; an end-effector can be thought of as the end point of a mass-spring chain whose position corresponds to the location of the most distal mass relative to the anchoring point in the vocal tract of the chain's most proximal spring.

For some gestures, it is the position of the end effector itself that captures the state of the system with regard to the given task. For example, the tongue tip position with respect to the (fixed)

[2]The alveolar ridge and the attachment point of the upper lip are specified within the vocal tract geometry and are immovable.

alveolar ridge is on its own sufficient to describe the degree of realization of an alveolar consonant. In the special case of the bilabial stop, however, there are two end effectors that contribute to a single occlusion. The positions of the two lip end effectors on their own do not provide enough information about the realization of the stop. Rather, it is a combination of both (the lip aperture) which is needed to assess whether the stop has been achieved. Such target relevant combinations of the end-effector positions are called *tract variables*. Formally, they can be defined through a functionally specified mapping projecting the end-effector positions into a relevant space capturing the degree of task completion, as in the standard task dynamic model.

We thus have a chain of concepts from gestures, through tract variables, and end effectors, to pure articulators. In Appendix 1 we describe the mappings between these various coordinate systems in greater detail. A complete technical account, including some implementation details not directly relevant to the present purposes, is provided in Simko (2009).

Gestures

As in Articulatory Phonology, we start with a gestural score, much like Figure 1. Activation functions are simple step functions, so that a gesture is either active or inactive. Looking ahead somewhat, one principal motivation for the current model is that we will ultimately be able to derive the details of the gestural score (the times of gesture on- and offsets, along with overall system stiffness) from a parametric cost function. In our simplified vocal tract, only a very few gestures are distinguishable, and they stand in one-to-one relation to phonetic segments. They are /b/, /d/, /i/ and /a/. Voicing is not modeled, so the system could not distinguish between /b/ and /p/, for example. This radical simplification is deliberate, and allows us to concentrate on those dynamical properties of an embodied system that constrain and shape its ultimate form of movement. Extension of the vocal tract to include more gestures can be undertaken once the basic sequencing principles are in place.

Tract Variables and End Effectors

There are three tract variables in our model:

- z_{TB} is the vertical position of the tongue body surface essential for the articulation of syllabic nuclei,
- z_{TT} is the vertical position of the tongue tip placement instrumental in forming an alveolar constriction, and
- z_{LA} is the distance between lips, the lip aperture, instrumental in bilabial stop realization.

The tongue body and tongue tip tract variables express the positions of their respective end effectors relative to a fixed reference point on the vertical axis, while z_{LA} simply contains an absolute distance between lips. The state of a functionally relevant overall constellation of the vocal tract is captured by the tract variable vector

$$\mathbf{z} = (z_{TB}, z_{TT}, z_{LA})^T.$$

Each gesture prescribes a target for the associated tract variable. Gestures /i/ and /a/ are associated with the variable z_{TB}, gesture /d/ with z_{TT} and gesture /b/ with z_{LA}. Numeric ranges

over which the tract variables operate are delimited by the physical boundaries of the oral cavity. The tongue body and tongue tip tract variables are restricted by the top of the mouth, and the lip aperture cannot be negative.

For the stop consonants these physiological boundaries act as the realization targets of associated gestures. The constriction target for gesture /d/ (for tract variable z_{TT}) is the position of the alveolar ridge, $z_{/d/}$. For bilabial /b/, it is zero distance between the lips, $z_{/b/} = 0$ (constriction target for z_{LA}).

The vocalic tongue body targets $z_{/i/}$ and $z_{/a/}$ must lie below the top of the mouth, and must reflect the relative height of modeled vowels, i.e.

$$z_{/a/} < z_{/i/} (< z_{/g/}),$$

where $z_{/g/}$ is a corresponding position on the velum (not yet implemented). The positions $z_{/i/}$ and $z_{/a/}$ act as constriction targets for tract variable z_{TB}.

In our model we consider four end effectors responsible for the constriction formations required for the realization of model gestures (see Figure 2). The end effector variables are

- the position of the tongue body surface point Z_{TB} relevant for our syllabic nuclei,
- the tongue tip placement Z_{TT} needed for the alveolar stop /d/, and
- the positions of the upper lip Z_{UL} and the lower lip Z_{LL}, instrumental in the bilabial stop /b/ production.

All four end effector variables capture vertical positions of the associated end effectors with respect to a fixed reference point. The *task matrix* is a 3×4 matrix mapping from the end effectors to tract variables (see Appendix 1).

End Effectors and Pure Articulators

The position of each end effector can be uniquely expressed as a function of the lengths of underlying pure articulators. In our model, we consider five pure articulators: the jaw, the tongue body, the tongue tip, and the lower and upper lip pure articulators. Their current states are captured by the following variables: y_J, y_{TB}, y_{TT}, y_{LL}, and y_{UL}.

Despite the fact that these variables share some of the subscripts with the tract and the end effector variables, there is an important difference[3]. With the exception of y_J and y_{UL}, each pure articulator variable allows calculation of the position of the associated end effector *relative* to the position of another articulator to which it is anatomically linked. So, y_{TB} is the distance between the vertical displacement of the tongue body end effector and the jaw, and is thus equivalent to the length of the tongue body itself; y_{TT} is the distance between the vertical displacements of the tongue tip and the tongue body end effectors, and is thus the length of the tongue tip; and y_{LL} is the distance between the lower lip end effector and the jaw. The variables y_J and y_{UL} contain vertical displacements of the jaw and the upper lip within a fixed spatial reference frame. Please note that in our model, positive values of the variables y_J and y_{UL} indicate *downward* displacement

[3]The only reason for this manner of indexing is aesthetic. In our opinion, the more accurate expressions, e.g. y_{TT-TB}, look cumbersome. As each pure articulator is functionally unambiguously attached to a single dominant articulator, one of the indices is always redundant.

of the jaw and the upper lip, while for the remaining variables they mean *upward* displacement relative to the associated dominant articulator.

The pure articulator vector

$$\mathbf{y} = (y_J, y_{TB}, y_{TT}, y_{LL}, y_{UL})^T$$

contains the lengths of all pure articulators at a given time.

A guiding principle within the present model is that the forces giving rise to the desired dynamic behavior be applied at the level of the pure articulators. The position of each end effector and the value of each tract variable depend on the values of several pure articulator variables. For example, the position of the tongue tip end effector Z_{TT} depends not only on the value of the tongue tip pure articulator y_{TT}, but also on the values of the tongue body and jaw pure articulator variables y_J and y_{TB} to which the tongue tip is anatomically linked. Table 2 shows the nature of these anatomical connections between our model pure articulators and the end effectors. The full circles are placed in the cross-sections of the rows (end effectors) and columns (pure articulators) for which there exists an anatomical link accounted for by our model.

	y_J	y_{TB}	y_{TT}	y_{UL}	y_{LL}
Z_{TB}	●	●	○	○	○
Z_{TT}	●	●	●	○	○
Z_{UL}	○	○	○	●	○
Z_{LL}	●	○	○	○	●

Table 2: Dependence of the end effector positions on the values of pure articulator variables. Full circles indicate a presumed (anatomical) connection, empty circles indicate no functionally relevant link.

When the vocal tract is in a resting state, i.e. not speaking, the values of all pure articulator variables are set to a resting value, e.g. 0. When the model is prepared for speaking, the articulators attain a different constellation tuned for the efficient execution of a (language-dependent) collection of speech gestures in its gestural repertoire. This speech-ready posture is determined by the speech-ready states of the pure articulators. We use a constant vector $\mathbf{y_0}$ to store these fixed speech-ready positions of model pure articulators.

Articulators in the human vocal tract exhibit complex linkages. A downward rotation of the jaw has a greater effect on the lip than on the tongue root, for example. Although our model represents only a single vertical dimension of end effector movement, some of these linkages can be incorporated by the mapping between pure articulator movement and the movement of end effectors. The sensitivity to the horizontal position along the front-back axis of the end effectors that affects the relative magnitudes of their vertical displacement can be captured by assigning different weights to the effect of pure articulator changes (e.g., jaw) on the positions of the linked end effectors. The distance by which the end effectors move increases, the further they lie from the jaw joint: $\Delta Z_{TB} < \Delta Z_{TT} < \Delta Z_{LL}$ for a single value of Δy_J. Relatedly, it is known from Gay (1977) that if the tongue tip is not engaged in the realization of an active gesture, vertical tongue body movements are generally accompanied by a movement of the tongue tip in the same direction but of a smaller magnitude. We therefore introduce the additional constraint that as the tongue body raises, i.e. the tongue body pure articulator variable y_{TB} increases, the tongue tip end effector moves in the same direction, but over a shorter distance: $\Delta Z_{TT} < \Delta y_{TB}$.

Together, these constraints can be built into a matrix that is used to map from tract variables through end effectors to pure articulators (Appendix 1). A mapping of this sort is required by Equation 2, but it is not provided explicitly in Saltzman and Munhall (1989). The anatomy matrix defined in Appendix 1 makes this mapping explicit and underwrites our related Equation 7 (see next section).

Embodied Task Dynamics: The Dynamics

With the basic model levels in place, we are in a position to define the dynamics. In contrast to conventional task dynamics, we need a *bidirectional* mapping from tract variables to articulators and vice versa.

We start with the definition of a time dependent restriction of the range of the anatomy matrix as defined above to only those tract variables which are active at a given time. This restriction is formally equivalent to the active-gesture gating principle of task dynamics. Then we proceed with the formal definition of the second order linear task dynamics acting on the active tract variables. Finally, we proceed with a projection of the tract variable task dynamics into the pure articulator space in order to obtain the articulator trajectories eliciting the required kinematics of the tract variables.

Task Dependent Anatomy Mapping

In mapping between the tract variable, end effector and pure articulator coordinate systems, we restrict our computations at all times to those variables relevant to the set of currently active gestures, and hence tract variables, using a restricted anatomy matrix, as described in Appendix 2. Our technique of defining the task dependent projection of pure articulator behavior to the subspace of active tract variables is equivalent to the gestural gating approach of task dynamics in Saltzman & Munhall (1989).

Note that a gestural score can, in principle, prescribe the co-production of multiple gestures posing conflicting targets on the *same* tract variable, for example, concurrent activity of vocalic gestures /a/ and /i/ driving the tongue body to two different position. This presents no problem in our treatment. According to our definition, the active anatomy matrix will contain two copies of the same row associated with the shared tract variable, and the active target vector will consist of the positions of conflicting targets. The resulting dynamics of the shared tract variable will correspond to an average gestural blending of the combined task dynamics.

Tract Variable Dynamics

In our model, active tract variables are driven towards their gestural targets in the manner of a simple mass-spring second order dynamical system. This treatment is in agreement not only with the task dynamic implementation of Articulatory Phonology, but also with more general theory of skilled target-oriented motor action (Kelso et al., 1986; Saltzman and Kelso, 1987).

If $\mathbf{z} = \mathbf{z}(t)$ is a vector containing the collection of all n tract variables *active*[4] at time t, and

[4]In the subsequent treatment, the variable $\mathbf{z}$ will represent the vector of all tract variables *active* at the given time, and not the vector of all tract variables defined in our model. Similarly, the vector $\mathbf{z}_0$ will contain the targets of all *active* gestures, and $\mathbf{A}$ will stand for the *active* anatomy matrix. We hope, that this convention introduced in order to avoid superficial indexing in our complex equations does not confuse the reader.

$\mathbf{z}_0 = \mathbf{z}_0(t)$ the corresponding vector of gestural targets, the tract variable dynamics is a solution of the following system of second order differential equations:

$$\mathbf{M}_z\ddot{\mathbf{z}} = -\mathbf{K}_z(\mathbf{z} - \mathbf{z}_0) - \mathbf{B}_z\dot{\mathbf{z}}. \tag{6}$$

The gestural stiffness matrix $\mathbf{K}_z$ is a diagonal matrix $\text{diag}(k_1, \ldots, k_n)$, where each k_i describes the gesture-dependent stiffness of the tract variable in achieving the target z_0. We presume that the task stiffness k_i is associated with the phonological *nature* of the given gesture: it is greater if the ith active tract variable corresponds to a consonantal gesture, and smaller for vocalic gestures. For parsimony's sake we presume that the numerical values of all gestural stiffness coefficients are analytically linked. Each stiffness coefficient k_i can be expressed as a fixed multiple of an overall system-wide stiffness parameter k, i.e., $k_i = \kappa_i k$ where κ_i is a scaling coefficient expressing the relative strength of the given gesture. Although the values of stiffness coefficients used in the realization of a given gestural score will vary depending on prosodic demands (e.g., speaking rate), they are constrained to vary in an orderly manner as functions of a single system parameter k.

As highlighted earlier, traditional task dynamics provides no interpretation for the task mass matrix $\mathbf{M}_z$ other than a formal expression of *abstract* task masses unrelated to the physical properties of any articulatory structures. In the task dynamic model, $\mathbf{M}_z$ is a redundant identity matrix, assigning unit loads to the abstract tract variables. At this point, however, we make no such assumptions about the nature of the task mass matrix. As we shall see below, this matrix can be redefined to reflect the mass distribution of pure articulators with regard to the active task, and, crucially, to provide a formal source of dynamical coupling among the tract variables reflecting the embodied nature of speech production.

In agreement with the task dynamic treatment of Articulatory Phonology, we presume the task damping $\mathbf{B}_z$ to be critical, i.e., $\mathbf{B}_z = 2\sqrt{\mathbf{M}_z \cdot \mathbf{K}_z}$. The damping coefficients in Equation 6 are thus analytically fully specified by the values of the stiffness and mass dynamical parameters. Therefore, the damping coefficients are *not* free dynamical parameters of the tract variable task dynamics considered here.

In order to model oral cavity boundaries, we expand the damping component of each equation by an extra expression that is a function of $|z - z_b|$, the distance between the tract variable z and the position of the physical boundary z_b relevant for the tract variable (e.g., the absolute position of the alveolar ridge in the end effector space for the tongue tip tract variable). This expression is incorporated into current simulations as defined in Equation 23 of Appendix 2.

Pure Articulator Dynamics

Equation 6 describes the task dynamics of our model speech production system at the level of tract variables, at which there is a one-to-one mapping between behavioral goals (gestures) and tract variables. The next step is to find an equivalent dynamical description in the underlying coordinate system of pure articulators, where the mapping between behavioral goals and articulators is potentially many-to-many. In other words, we need to transform Equation 6 into the space of pure articulators $\mathbf{y}$ so that the resulting pure articulator kinematics projected to the tract variable system $\mathbf{z}$ through the active anatomy mapping $\mathbf{A}$ is the same as that determined by Equation 6.

We use the methodology introduced in the task dynamic model of Articulatory Phonology (Saltzman and Munhall, 1989). We first express the mapping from articulators to tract variables for position, velocity, and acceleration (Equations 7–9) using the anatomy matrix $\mathbf{A}$. We then

invert the forward acceleration mapping by deriving and using a pseudo-inverse of the Jacobian of this transformation (cf. Equation 5).

The linear active anatomy transformation expresses the relation between the currently active tract variables and the set of model pure articulators. This is expressed in matrix form as

$$\mathbf{z} = \mathbf{A}\mathbf{y}. \tag{7}$$

Since the elements of $\mathbf{A}$ are constant, the Jacobian transformation matrix of this mapping is simply the matrix $\mathbf{A}$, and its time-derivative is $\mathbf{0}$, so the following relationships hold for the time-derivatives of $\mathbf{z}$:

$$\dot{\mathbf{z}} = \mathbf{A}\dot{\mathbf{y}}, \tag{8}$$

$$\ddot{\mathbf{z}} = \mathbf{A}\ddot{\mathbf{y}}. \tag{9}$$

Substituting into Equation 6 we get a system of second order differential equations expressed in terms of pure articulators and providing a possible dynamical description of pure articulator behavior yielding the desired dynamics at the tract variable level:

$$\ddot{\mathbf{y}} = \mathbf{A}^{*\mathbf{W}}\mathbf{M}_z^{-1}[-\mathbf{K}_z(\mathbf{A}\mathbf{y} - \mathbf{z}_0) - \mathbf{B}_z\mathbf{A}\dot{\mathbf{y}}], \tag{10}$$

where $\mathbf{A}^{*\mathbf{W}} = \mathbf{W}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T)^{-1}$ is a weighted pseudoinverse of the anatomy matrix $\mathbf{A}$ for a suitable 5×5 weight distribution matrix $\mathbf{W}$, i.e., $\mathbf{A}\mathbf{A}^{*\mathbf{W}} = \mathbf{I}$ (Klein and Huang, 1983). The weight matrix, $\mathbf{W}$ is not fully determined, and may take several different forms.

As required, a transformed solution of differential equations 10 yields the tract variable trajectories specified by Equation 6. Indeed, if we multiply both sides of Equation 10 by the active anatomy matrix $\mathbf{A}$ and reverse the substitution 7–9 we get a system of equations mathematically equivalent to Equation 6.

Equation 10 is derived in a way equivalent to the derivation of Equation 5 used by the task dynamic implementation of Articulatory Phonology for computing the behavior of model articulators (Saltzman and Munhall, 1989). The notable difference between these two equations is the presence of the task mass matrix $\mathbf{M}_z$ in Equation 10. If we were to follow the example of the original approach and replace this matrix with the identity matrix, that would lead to the following solution:

$$\ddot{\mathbf{y}} = \mathbf{A}^{*\mathbf{W}}[-\mathbf{K}_z(\mathbf{A}\mathbf{y} - \mathbf{z}_0) - \mathbf{B}_z\mathbf{A}\dot{\mathbf{y}}]. \tag{10a}$$

As argued in the section "Task Dynamics with Abstract Tasks", the pure articulators behave as genuinely force driven dynamical components if, and only if, their behavior depends on the distribution of masses among the articulators in a direct manner, regardless of the set of active tasks driving the system's behavior. We also pointed out that the weight matrix used in the definition of the pseudo-inverse of the anatomy matrix in Equation 10 can be used, to some extent, to account for mass distribution among the pure articulators of the vocal tract. However, as the following simple thought experiment shows, this method is not universally suitable to embody the task driven articulatory system.

Let us extend the set of tasks defined for our simplified vocal tract by a (presumably non-linguistic) task imposing a target on the position of the jaw. Technically, this task is defined by

adding an extra row (1 0 0 0 0) to the anatomy matrix $\mathbf{A}$ (Equation 17, Appendix 1). If this task is activated concurrently with a vocalic task imposing a target on the tongue body and an alveolar stop task /d/ imposing a target on the tongue tip, the mapping projecting the positions of pure articulators of the tongue-jaw sub-structure of the model vocal tract to the active tract variable values becomes non-redundant. In other words, there is a single constellation of the jaw and tongue pure articulator positions leading to a successful realization of these tasks.

Importantly, in this case, the active anatomy matrix $\mathbf{A}$ effectively reduces to an invertible sub-matrix within which all non-zero elements are contained. The weight matrix $\mathbf{W}$ used in Equation 10a thus plays no role whatsoever in determining the behavior of pure articulators, i.e., the equation does not refer to the articulator masses at all. Therefore, the task-driven kinematics of the pure articulators obtained as solutions of Equation 10a cannot be regarded as the result of embodied force-driven dynamics.

The question we need to answer is whether there exists a re-interpretation of the task mass matrix $\mathbf{M}_z$ which would allow us to derive a form of Equation 10 that satisfies the requirement of genuinely embodied articulatory behavior formulated above. As we show in detail in Appendix 2 the answer to this question is affirmative. We summarize the derivation briefly here:

First, we define a 5×5 diagonal matrix $\mathbf{M}_y$ describing the distribution of masses acted upon by the pure articulators. In the spirit of mass-spring dynamical systems, matrix $\mathbf{M}_y$ contains the masses associated with the pure articulators on its diagonal, and zeros elsewhere.

If the task mass matrix, $\mathbf{M}_z$, is set to be related to this pure articulator mass matrix, $\mathbf{M}_y$, in the following way

$$\mathbf{M}_z = \mathbf{A}\mathbf{M}_y\mathbf{A}^{*\mathbf{M}_y}, \tag{11}$$

then Equation 10 can be expressed[5] as

$$\mathbf{M}_y\ddot{\mathbf{y}} = \mathbf{A}^{*\mathbf{I}}[-\mathbf{K}_z(\mathbf{A}\mathbf{y} - \mathbf{z}_0) - \mathbf{B}_z\mathbf{A}\dot{\mathbf{y}}]. \tag{12}$$

Note that the right hand side of Equation 12 is again, as in Equation 10, equal to the right hand side of the task dynamic differential Equation 6 recast to the pure articulator system using a pseudo-inverse of the active anatomy matrix $\mathbf{A}$.

As the influence of masses on the task-oriented behavior of the pure articulators is represented by the mass matrix $\mathbf{M}_y$ in Equation 12, we do not need to impose an additional scaling of the pseudo-inverse using a weight matrix related to articulator masses. Therefore, for the sake of generality, we set the weight matrix to be the identity matrix.

Unlike the kinematic version (Equation 10a) used by the traditional task dynamic implementation, Equation 12 characterizes genuine *dynamic* behavior of the model pure articulators dependent on the distribution of masses upon which the articulators act. Indeed, the influence of the masses on the pure articulator dynamics is cleanly separated from the effect of concurrently active tasks (the masses do not feature on the right hand side of Equation 12 at all). This influence is thus task independent and remains the same even in the case of a non-redundant system, as discussed earlier. Moreover, as the pure articulator mass matrix, $\mathbf{M}_y$, is diagonal, Equation 12 allows a direct expression of the force $\vec{F}_i = m_i\ddot{y}_i$ acting on each individual pure articulator.

A significant departure from the traditional task dynamic (and Articulatory Phonology) approach introduced by linking the dynamics of articulators and tract variables in a bi-directional

[5]Note that *both* Equations 10a and 12 are mathematically equivalent to the general description of pure articulator dynamics captured by Equation 10. They, however, crucially differ in conceptualization of the task mass component $\mathbf{M}_z$.

fashion is that there is a resultant coupling, not only among the articulators, but among the tasks as well. The task mass matrix, $\mathbf{M}_z$, specified by Equation 11 is generally not diagonal. It contains values off its main diagonal if the concurrently active gestures involve overlapping sets of pure articulators. This non-diagonality introduces a dynamical coupling among tract variables. The coupling depends on the nature of the concurrently active tasks and the anatomy of the model vocal tract.

The dynamic behavior of a tract variable when active on its own is, therefore, not necessarily the same as its behavior when performed alongside another task involving some of the pure articulators engaged by the given tract variable. In contrast, the original task dynamic implementation presumes an uncoupled nature of the dynamics of the pre-tuned individual tasks. The system's context sensitivity is manifested exclusively at the model articulator level.

Our interpretation introduces an additional constraint at the task level that makes the manner of their realization sensitive to the underlying architecture of the articulatory components. We believe that this approach better reflects the embodied nature of skilled motor behavior. Moreover, as we argue in the next section, it allows us to meaningfully formalize the notion of articulatory effort that enables our optimization approach to gestural sequencing.

Speech-Ready Dynamics

When the vocal tract organizes itself for speech action, the articulators adopt positions suitable for reaching possible targets imposed on them by incoming tasks. These positions reflect the readiness of the oral cavity components to be used for speaking as opposed to some other action in which they may be involved, for example chewing, or swallowing. The articulators return back to these speech ready positions when they are not engaged in any active task.

We model this influence on the vocal tract articulators by including an additional component influencing the behavior of pure articulators. This additional dynamical component represents restoring forces on each pure articulator separately in a task independent fashion, much as the neutral attractor within traditional task dynamics (Saltzman and Munhall, 1989).

Again, we model the effect of this dynamical component on each pure articulator using critically damped mass-spring dynamics. Formally, the speech ready dynamics can be expressed as:

$$\mathbf{M}_y\ddot{\mathbf{y}} = -\mathbf{K}_{y0}(\mathbf{y} - \mathbf{y}_0) - \mathbf{B}_{y0}\dot{\mathbf{y}} \tag{13}$$

where $\mathbf{M}_y$ is the same pure articulator mass matrix used in the previous section, $\mathbf{K}_{y0}$ is a diagonal matrix with speech ready stiffness coefficients of individual pure articulators on its diagonal, and $\mathbf{B}_{y0} = 2\sqrt{\mathbf{M}_y \cdot \mathbf{K}_{y0}}$, ensuring critical damping of the speech ready dynamics of each pure articulator. All of these matrices are diagonal, so the speech ready dynamics is uncoupled. Just as with the task stiffness coefficients, the speech ready stiffness coefficients are scaled by the overall system-wide stiffness scaling factor, k.

The speech ready dynamics plays a dual role. Firstly, it guarantees that the articulators return to their initial positions when there are no active tasks influencing their behavior. Importantly, within our optimization paradigm discussed in the remainder of this paper, it also acts as an incentive to disengage the tasks when their targets are sufficiently met, by making explicit the force keeping the articulators away from their respective speech ready positions. Therefore, the speech ready dynamics is always active, and its influence on pure articulator kinematics is added to that of the task dynamics expressed in Equation 12.

If $\ddot{\mathbf{y}}_{task}$ is the acceleration imposed on pure articulators by task dynamics (Equation 12) and $\ddot{\mathbf{y}}_{sr}$ by the speech ready dynamics (Equation 13), then the overall acceleration of pure articulators is:

$$\ddot{\mathbf{y}} = \ddot{\mathbf{y}}_{task} + \ddot{\mathbf{y}}_{sr}.$$

The force exerted by the task goal and the force exerted by the speech ready dynamics act antagonistically. As a result, the task targets need to be adjusted somewhat to ensure that the resulting equilibrium position satisfies the gestural goal. For stop consonants, this is easily achieved by placing the target a few millimeters behind the point of contact. For vowels, the requisite adjustment to the target has hitherto been determined by trial and error. Details of parameters used are provided in Appendix 3.

Optimization

So far, we have been able to provide a re-formulation of task dynamics with the innovation that tasks are now embodied in the articulators, in the sense that the dynamical behavior of the tasks is critically dependent on the physical properties of the articulators that ultimately achieve the task goals. The result is that the inertial properties of the articulators constrain the evolution over time of tract variables as well as articulators, and there is the additional side-effect that tasks, as well as articulators, now exhibit mutual coupling. This goes some way towards addressing one perceived shortcoming of the original formulation of task dynamics as it might apply to speech: the disembodied, abstract, nature of mass within the system. However, our vocal tract geometry is highly simplified and stylized at this point, and the effort would scarcely seem worthwhile, were it not for the opportunity now to address the second desideratum of an embodied task dynamics: We are now in a position to exploit the system to constrain the temporal sequencing of gestures, providing a way of exploring inter-gestural coupling, and defining a space within which sequencing constraints can be expressed in a principled way, and their consequences within the model studied. With that comes the prospect of a rich interaction between model development and empirical data-driven inquiry in the future.

A physically embodied system incurs a metabolic cost when it moves. Prohibitively high costs rule out most movements, and a key insight in the study of motor behavior is that well-practiced skilled movements are usually interpretable as optimal with respect to suitably defined constraints, such as effort. For example, in a famous study, Hoyt and Taylor (1981) showed that horses naturally use gaits that minimize the metabolic cost required to travel a given distance. The cost function of travelling speed (expressed as oxygen consumption per unit distance travelled) formed U-shaped curves with distinct minima, one for each gait (sequencing pattern). These minima coincided with the speeds that horses spontaneously adopt when moving freely. This suggests that the sequencing details of locomotion are fine-tuned (presumably by evolution) in order to minimize, among other variables, the metabolic cost.

Of course the simplest way to minimize energy expenditure is to do nothing. We therefore presume that a communicative system places a premium on being understood, which in the current context means that energetic costs need to be traded off against the costs associated with producing faulty or imprecise articulation. The idea that speech patterns can be usefully understood as representing a compromise between the twin demands of precise articulation and moving as little as possible is not new. It forms the core of Lindblom's well known Hyper-Hypo continuum (Lindblom, 1990) and motivates much of the theory of Emergent Phonology (Lindblom, 1999). Keller

(1987) has expressed the alternative opinion that energetic considerations are unlikely to play an explanatory role in understanding speech movement, because of the relatively slight masses and the correspondingly powerful muscles involved. The present model provides an opportunity to assess the potential role of such constraints in speech movement.

We augment the twin and competing constraints of minimizing articulatory effort and maximizing intelligibility with a third factor, utterance duration. There are several reasons to motivate this choice. Firstly, articulation rate, which manifestly influences durations, is largely independent of the Hypo-Hyper continuum. Gay (1981) reported that changes in speaking rate do not necessarily lead to the consequences implied by H&H Theory alone. People can speak quickly without undershooting articulatory targets, and slowly with imprecisely realized underlying gestures. In addition to adjustments of segmental duration and articulatory displacement, changes in speaking rate can be elicited by means of non-linear alterations to articulatory velocity and to intrasyllabic coarticulation.

Secondly, including duration within a parametric cost function allows us to investigate the consistency of inter-gestural phasing as a function of speech rate. It is well known that some temporal features of speech are highly malleable, while others are relatively insensitive to rate change, and that this dependence on rate is highly individual (Gay, 1981).

Finally, because of the role played by masses in our system, it is possible to model inter-individual differences arising from anatomical and physiological variation, though much of this work remains to be done. The degree to which duration is independent of articulatory precision, for example, may exhibit large inter-individual variation.

Quantifying Articulatory Effort

We presume that the metabolic cost associated with a specific speech movement, or a sequence of movements, is linked to the magnitude of forces acting on the vocal tract articulators. Within our modeling framework, we identified the pure articulators as the effectors directly influenced by the muscle action underlying speech production. The embodied version of task dynamics, presented in this paper, allows us to evaluate the forces acting on the pure articulators in a straightforward manner, in order to achieve the tasks prescribed by the gestural score and the restoring forces imposed by speech-ready dynamics of the system.

We define the articulatory effort E as the integral of the magnitudes of all these forces acting on each individual pure articulator during the realization of a given gestural score. The only force ignored in this evaluation is the damping force invoked when the end effectors hit the physical boundaries of the oral cavity. We presume that this passive force is linked to the elasticity of the tongue and lip tissue rather actively generated in order to produce a prescribed utterance and therefore it does not incur a metabolic cost.

Formally, if $\ddot{y}_{i,task}$ and $\ddot{y}_{i,sr}$ are the task dynamics and speech ready dynamics accelerations imposed on ith pure articulator by the dynamical systems defined in Equations 10 and 13, respectively, and m_i is the mass acted upon by the articulator, the magnitude of the sum of active and restoring forces acting on this articulator at any given moment is

$$F_i = |m_i\, \ddot{y}_{i,task}| + |m_i\, \ddot{y}_{i,sr}|.$$

The articulatory effort is then defined as

$$E = \sum_i \int_{T_b}^{T_e} F_i \, dt,$$

where the sum ranges over all pure articulators of the system and T_b and T_e are the onset of the activation of the first gesture and the offset of the activation of the last gesture in the gestural score, respectively.

The Parametric Cost Function

We now define a cost function with three components:

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

where α_E, α_P and α_D are simple scalar weight coefficients. Articulatory effort (E) is calculated as above. The second term is listener-oriented, and relates to the extent to which gestures achieve their goals, thus, presumably producing intelligible speech. This 'parsing' cost (P) seeks to reward gestures that reach or approximate their targets, and penalize undershoot or overly lax articulation. Target approximation means slightly different things for vowels and consonants, respectively.

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

$$p_v(t) = 1 - \left| \frac{z'_v - z(t)}{z'_v - z_0} \right|.$$

For each vowel gesture v, the estimate p_v is thus a time function depicting the level of achievement of the gesture's target.

For consonants, target approximation is not sufficient, and we define a simple binary function: If closure has been achieved at time t, $p_c(t) = 1$. Otherwise $p_c(t) = 0$.

The cost of parsing the given gesture g by the listener (whether g is a vowel or a consonant) depends also on the duration of the time interval during which the gesture is realized with sufficient precision, i.e. during which the precision estimate $p_g(t)$ exceeds a given threshold. The longer the gesture is articulated, the easier it is for the listener to recognise it. This intuition is captured by the duration estimate function, $d_g(t)$ which increases rapidly and monotonically to asymptote during the interval of the gesture's prominence. The parsing cost P_g for each gesture is the sequence is then expressed as a combination of these two estimates

$$P_g(t) = \max_t p_g(t) \max_t d_g(t) \tag{14}$$

where the maxima are taken from the interval of sufficient prominence of the given gesture g. A gesture is considered to be sufficiently prominent if the closure is achieved (for consonants) or if the associated tract variable is close to its gestural target (for vowels; $p_g > 0.8$). Moreover, in order to be prominent, the vocalic gesture must not be occluded by a concurrently active consonantal gesture, that is, the vowels are not realized during the overlapping consonantal closure even when their articulatory target is sufficiently approached. Similarly, the consonant with more frontal place of realization, e.g. /b/, occludes a consonant realized further back in the vocal tract, i.e., /d/. This

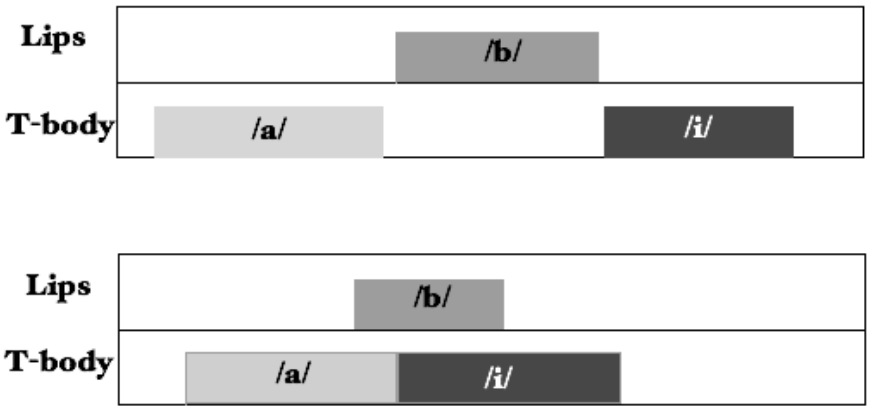

Figure 3: Gestural score before and after optimization.

limits the degree of consonantal overlap that can occur in an optimal score. The parsing cost associated with the entire sequence is then a simple sum of all partial parsing cost elements P_g computed using Equation 14. Full details are provided in Simko (2009).

Finally, the Duration Cost (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 prominence interval (the point at which the final gesture is judged to be finished) in the realized sequence.

Together these three components, with their associated weights, allow us to attach a cost to any given realization of an intended sequence of gestures. This in turn means that we can search for optimal sequences. Figure 3 illustrates the optimization procedure we are aiming at. In the top panel, the gestural sequence /abi/ is specified prior to optimization. While the sequential order of gestures is respected, no relative timing relations among gestures are presumed. In the bottom panel, gesture activation length and the relative timing of gestures have emerged after the optimization process.

The optimization procedure searches the space of gestural activation patterns and pure articulator stiffnesses for global minima with respect to the given cost function. That is, we incrementally modify system stiffness, k, and the activation interval onsets and offsets of all gestures until an optimal configuration is found. We assume that our starting constellation (Figure 3, top) is non-optimal with respect to our chosen cost function. The cost for the starting configuration is computed, and then the optimization process is employed to perform gradient descent on the cost function, until a local minimum is reached[6]. If the local minimum proves stable with respect to several local perturbations, it is deemed optimal, and this provides us with our final gestural score, and system stiffness.

The following table presents the influences of some high level properties of the gestural score

[6]The optimization algorithm based on simulated annealing is used to identify the gestural sequences that are optimal with respect to the defined cost function. The objective function maps the activation onsets and offsets, plus the overall stiffness values to the overall cost value. Gradient descent on the objective function is then employed until the process reaches a local minimum. At that point, the variable values are randomly perturbed and the gradient descent continues. This process continues until the gradient descent fails to find a new local minimum after a given number of perturbations.

and system stiffness constellation on the constituent cost functions.

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

Table 3: The relation between changes in system stiffness and gesture activation length on the one hand and the three cost constituents on the other. Up-arrows mean increases, down-arrows refer to decreases.

Modeling Results

In assessing the performance of the model, it is important to bear in mind that it is not proposed as a control algorithm for on-line generation of motion in time. Rather, the use of optimization allows us to explore the space of possible gesture sequences with their associated costs, and to identify some as more efficient than others, in a precise sense. In this way, we seek to account for the form of motion observed, but remain agnostic as to mechanisms.

One of the first results obtained with the embodied task dynamic model is the simple fact that the optimization procedure converges, and that the resulting movements do not appear to violate any obvious intuitions or known detail about gestural sequencing. This may appear trivial, but it needs to be emphasized that the sequences obtained are fully automatic and result from gradient descent based on the above cost function and nothing else. Moreover, by adjusting the duration cost, it is possible to examine gestural sequencing at a range of rates, and in general we find a pronounced stability of relative timing. That is, the timing of one gesture expressed with respect to another is relatively stable across a range of rates.

In all our simulations of VCV and VCCV sequences, vowels and consonant activation overlap to a great extent, such that there is one sequence of vowel activations, and a parallel, but distinct sequence of consonantal activations. This separation is a familiar characteristic of phonological representations, where it is referred to as the separation of the vocalic and consonantal tier; it is also well documented in the phonetics literature (Browman and Goldstein, 1990). In the cost efficient sequences, syllabic nuclei are produced as a continuous sequence of (vowel) gestural activations interleaved with consonantal gestures. This phenomenon is stable over a range of speaking rates and is an outcome of the cost optimization and is not encoded as an explicit phonological rule.

Figure 4 shows optimized gestural scores and associated articulator traces for two utterances: /abi/ and /iba/. The gestural score is shown on top, and the movement traces for the tongue body (solid line) and upper and lower lips (dashed lines). The trace for jaw movement (lighter solid line at the bottom) is also shown. The vertical lines demarcate the period of consonantal closure. Lip movement thereafter is due to soft body compression, as the target for each lip is slightly beyond the point at which closure occurs (Löfqvist and Gracco, 1997). Tongue movement is smooth and continuous during consonantal closure.

For /abi/, tongue movement towards the second vowel precedes the point of consonantal closure, and the onset of lip movement towards the closure starts *before* the tongue movement. This remains robust at a range of rates (not shown), although for faster simulated utterances tongue movement starts relatively earlier during the bilabial gesture activation than for slower ones. On the other

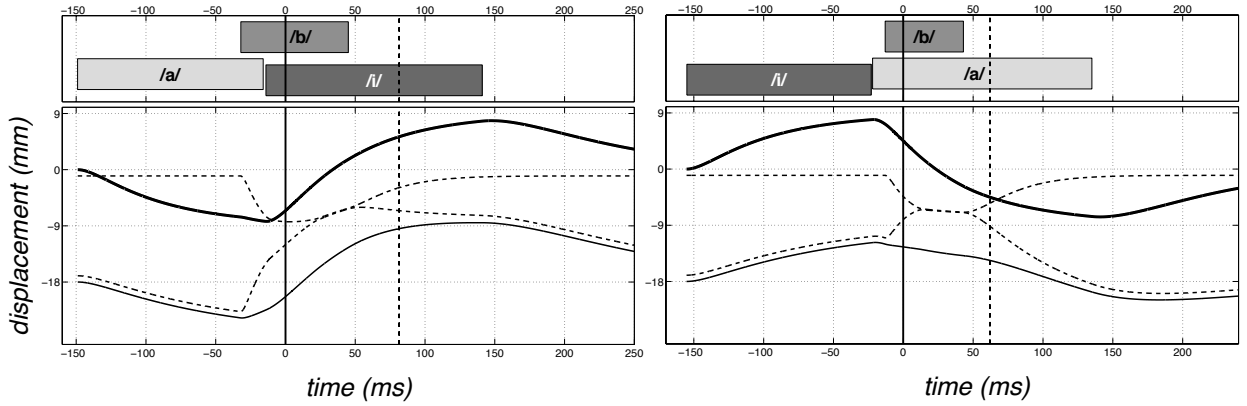

Figure 4: Articulator traces for utterances /abi/ and /iba/. Bold solid line = tongue body; Dashed lines = lips. Thin solid line = jaw.

hand, for /iba/, lip movement towards the closure starts slightly *later* than tongue movement toward the second vowel. These differences in gestural phasing which emerge in our simulations capture some of the variation that is best understood by consideration of the segment identity, and thus move beyond the gesture-independent phasing principles postulated in early accounts of AP.

Consider the data from Löfqvist & Gracco (1999) shown in Figure 5, which provides the relative timing of the onset of a consonantal bilabial gesture with respect to the intervocalic switch realized by the tongue body. If we restrict our attention to the asymmetric sequences with /a/ and /i/ vowels, 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 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/.

The gestural scores for sequences /abi/ and /iba/ plotted in Figure 4 tentatively capture these patterns. For /abi/, there is first the bilabial gesture followed by the transition between vocalic gestures followed by the lip closure achievement. For /iba/, the intervocalic switch leads the onset of the bilabial gesture and also the closure onset.

In describing the relative alignment of two gestures, a convention adopted by researchers within Articulatory Phonology (e.g. Browman & Goldstein, 1995) has been to use phase rather than clock time to describe when onsets, offsets, closures, releases, etc happen. To do this, one gesture is treated as a temporal referent for another. Each gesture has an associated second order mass spring dynamical system. If we neglect the critical damping term, this provides a periodic referent, and events can be described as occurring at specific phases of this underlying abstract cycle. Many outstanding questions about the relative timing of events can be couched in terms of phase invariance, phase variability, etc. Figure 6 illustrates how the abstract underlying cycle of a vowel may be used to index events such as the onset of a consonantal activation interval. In this way, one may also talk, for example, of the phase of consonant closure, with respect to the consonantal cycle itself. Phase values used in this way may also lie outside the range of [0,360] degrees, as the abstract underlying cycle can repeat to an arbitrary extent in either direction. The period of the

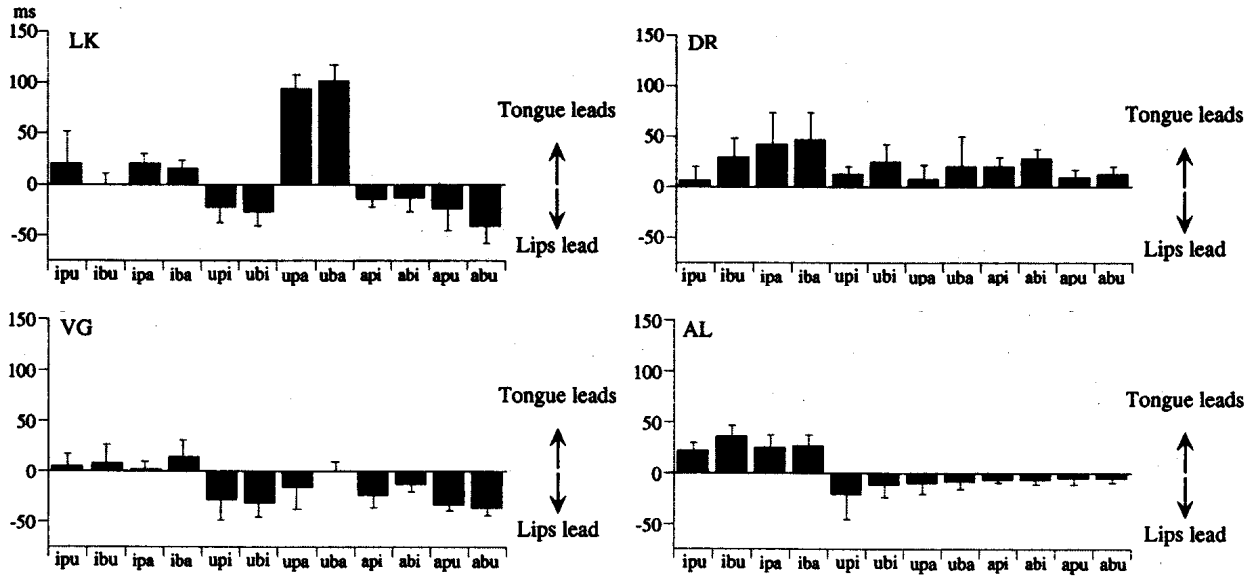

Figure 5: Interval between the onset of the tongue movement and the onset of the lip closing movement for the consonant, from (Löfqvist and Gracco, 1999). Standard deviations are also plotted.

underlying cycle is determined by the mass and stiffness of the articulator (not by the duration of the activation interval).

Much discussion within the Articulatory Phonology/Task Dynamics literature has centered on whether invariant phase relations are to be found among classes of gestures, e.g. between syllable initial consonants and following vowels (Browman and Goldstein, 1990). Such invariance would represent an important link between the regularity sought by phonological theory and its phonetic instantiation. The stringent notion of phase invariance was tentatively suggested in Browman & Goldstein (1990). This highly constrained view has more recently been relaxed and elaborated upon in both the phase window approach of Byrd (Byrd, 1996) and in the use of planning oscillators (Saltzman et al., 2008) as additional model components. Our modeling set up allows investigation of the relationship between optimal phasing among gestures and parameters such as speaking rate (through manipulation of the duration cost), and speaking style (through variation in the relative weight assigned to effort and parsing).

We can present initial results that examine the constancy of phasing as we vary the segments in a simple V_1CV_2 sequence ($V_1 \neq V_2$), . These are presented in Table 4. The first two data columns show the phase at which the C-closure and the V-V transition occur with respect to the underlying consonantal cycle. The following 4 columns change the temporal referent, and provide the phase of consonantal onset and closure expressed with respect to the first (data cols 3 and 4) and second (data cols 5 and 6) vowel, respectively.

These preliminary data from our simulations suggest, unsurprisingly, that some phase relations may be more constant than others. The point within the cycle of the first vowel at which consonantal closure occurs, for example, is much less variable than the point within the consonantal cycle itself at which closure occurs. As illustrated by these data, phasing details within an embodied production

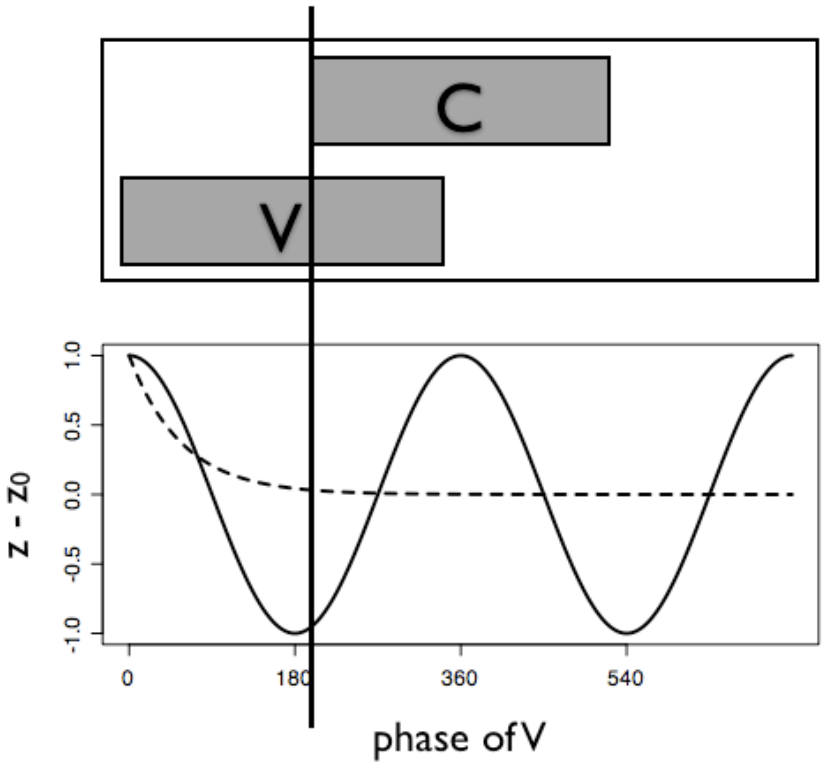

Figure 6: The top panel shows two activation intervals, one for a vowel and one for a consonant. In the lower panel, the dashed line illustrates the motion of the Tongue Body Constriction Degree tract variable associated with the vowel. Its change over time is critically damped. The solid line shows the evolution of the phase of a corresponding *undamped* oscillation associated with the vocalic gesture. The onset of the consonantal activation is thus at a phase of approximately 200° of the associated vowel cycle.

system are unlikely to exhibit invariant phase relations. The entangled influences of articulatory, functional, perceptual and efficiency constraints will inevitably lead to variation. The embodied modeling platform allows the exploration of phasing relations as a function of all these influences, and thereby opens up a novel and important field in which modeling and empirical inquiry can advance together. A fuller account of phasing variability, and the relation between our model and the phase window approach of Byrd (1996) will be provided in a subsequent article (Simko and Cummins, 2010).

Discussion

The focus of this article has been to present, in some detail, a modified form of the task dynamic framework in which tasks are realized in an embodied fashion. Two main motivations were provided: the intuitions of phoneticians that real masses and inertial properties are critical for a full description of phonetic gestures, and the long outstanding problem of uncovering sequencing principles that

Referent	Consonant		Vowel 1		Vowel 2	
	C closure	V-V switch	C onset	C closure	C onset	C closure
/abi/	226°	132°	246°	314°	−38°	30°
/adi/	179°	115°	252°	306°	−33°	21°
/iba/	93°	−56°	300°	328°	18°	46°
/ida/	84°	−65°	303°	328°	20°	45°

Table 4: The phases at which salient events happen expressed with respect to the underlying consonantal, initial vowel and final vowel cycles.

are capable of specifying the fine details of the temporal structure of interleaved and overlapping sequences of gestures.

While most of our efforts herein have been directed at fleshing out the details of an embodied task dynamics, we hope that sufficient detail has been provided that the reader can see how this enables the application of optimization principles to the sequencing problem in a principled fashion. We have presented only exemplary results herein. In a subsequent article, we discuss the many details involved in optimization, and the associated implications for sequencing patterns that result (Simko and Cummins, 2010).

As this work progresses, it is evident that theoretical choices can only be reasonably guided through empirical investigation of articulatory movement at a range of rates. The model has now been developed to a stage where a rich back-and-forth with empirical work is both possible and desirable. The parametric cost function developed here allows investigation of the relative importance of articulation rate, articulatory precision and effort in co-determining the phasing relations observed between gestures within and across tiers.

Even at this stage, the model would benefit from rich articulatory data obtained for a variety of syllables at a wide range of rates. Thereafter, several obvious avenues of exploration suggest themselves. The minimal vocal tract geometry we have employed needs to be enlarged. Addition of velar and glottal gestures seem to pose no substantial problem. Elaboration of the vowel space to include two dimensional movement is somewhat more challenging, but again should be possible in principle.

Beyond gestural modeling, there remains a host of questions about the relationship between timing at the gestural level and at higher levels, e.g. the phonological word, foot, and phrase, that need to be addressed. Most phonological theories have been relatively unencumbered by physical details to date. With the model we are developing here, it is to be hoped that an embodied and performative account of gestural timing may ultimately inform and shed some light on the relationship between temporal patterns observed at the gestural scale and above.

There are many more reasons to seek a thoroughly embodied account of skilled action, whether it be a speaker wrapping his tongue figuratively around the sonnets of John Donne, or a cockroach skillfully negotiating the varied landscape of a poet's kitchen. In each case, smooth, context-sensitive movement results from over-arching and context-independent behavioral goals. In the latter case, but perhaps in the former also, an understanding of complex action can not be couched in terms of an abstract and all-powerful brain or controller. Rather, smartness is built into the system, through evolution for the most part, but with some role for individual developmental history. Smartness is distributed throughout the organism, and inheres in the manner in which

physically real effectors coordinate with each other and with their environments. The elusive concept of "smartness" can find a more precise expression in our models through the definition of objective optimization functions that describe the resulting, efficient landscape of action and potential action. The role of the nervous system in a thoroughly embodied system is somewhat different from the detached controller of many cognitive models. Rather than issuing commands, the nervous system is one part of the machinery necessary to constrain a system of almost unlimited potential into a highly constrained goal-directed system whose several parts cooperate in achieving those goals.

References

Browman, C. and Goldstein, L. (1990). 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. Cambridge University Press, Cambridge.

Browman, C. and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica*, 49(3-4):155.

Browman, C. P. and Goldstein, L. (1995). Dynamics and articulatory phonology. In Port, R. F. and van Gelder, T., editors, *Mind as Motion*, chapter 7, pages 175–193. MIT Press, Cambridge, MA.

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

Fowler, C. A., Rubin, P., Remez, R., and Turvey, M. (1980). 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. (1977). Articulatory movements in VCV sequences. *The Journal of the Acoustical Society of America*, 62(1):183.

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

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

Hawkins, S. (1992). An introduction to task dynamics. In Docherty, G. J. and Ladd, D. R., editors, *Gesture, Segment, Prosody: Papers in Laboratory Phonology II*, pages 9–25. Cambridge University Press, Cambridge.

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

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

Kelso, J. A. S., Saltzman, E., and Tuller, B. (1986). The dynamical perspective in speech production: Data and theory. *Journal of Phonetics*, 14:29–60.

Klein, C. and Huang, C. (1983). Review of pseudoinverse control for use with kinematically redundant manipulators. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:245–250.

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, Dordrecht.

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

Löfqvist, A. and Gracco, V. (1997). Lip and jaw kinematics in bilabial stop consonant production. *Journal of speech, language, and hearing research: JSLHR*, 40(4):877.

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

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. (1991). The task dynamic model in speech production. In Peters, H. F. M., Hulstijn, W., and Starkweather, C. W., editors, *Speech Motor Control and Stuttering*, chapter 3. Elsevier Science.

Saltzman, E. (1995). Dynamics and coordinate systems in skilled sensorimotor activity. In Port, R. F. and van Gelder, T., editors, *Mind as Motion*, chapter 6, pages 149–173. MIT Press, Cambridge, MA.

Saltzman, E. and Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19(4):499–526.

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

Saltzman, E. and Munhall, K. (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 *Proceedings of the 4th International Conference on Speech Prosody. Brazil: Campinas.*

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. (2009). Sequencing of articulatory gestures using cost optimization. In *Proceedings of INTERSPEECH 2009*, Brighton, U.K.

Simko, J. and Cummins, F. (2010). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science*. Submitted.

Acknowledgements

The present work forms part of the Ph.D. thesis of the first author. It was funded through a Science Foundation Ireland Principal Investigator grant, 04/IN3/I568, to the second author. A large debt of gratitude is due to Elliot Saltzman whose persistent engagement and criticism have helped to substantially improve and refine this model, and to clarify many issues for us. Thanks are also due to Louis Goldstein, Dani Byrd, Hosung Nam and Michael O'Dell for very helpful discussions and suggestions. We would also like to thank the editor and three anonymous reviewers, who have helped greatly improve this manuscript.

Appendix 1: Model Architecture and Mappings

The temporal information in a gestural score can be encapsulated in an activation time vector

$$\mathbf{a}(t) = (a_{/i/}(t), a_{/a/}(t), a_{/b/}(t), a_{/d/}(t))^T.$$

The target vector

$$\mathbf{z}_0 = (z_{/i/}(t), z_{/a/}(t), z_{/b/}(t), z_{/d/}(t))^T,$$

related to the activation vector $\mathbf{a}(t)$, encapsulates the numerical values of all gestural targets considered in our system.
We define the end effector variable vector

$$\mathbf{Z} = (Z_{TB}, Z_{TT}, Z_{UL}, Z_{LL})^T$$

The tract variables are linked to these end effectors in a straightforward, linear fashion:

$$\begin{aligned} z_{TB} &= Z_{TB}, \\ z_{TT} &= Z_{TT}, \\ z_{LA} &= Z_{UL} - Z_{LL}. \end{aligned}$$

Formally, this linear relationship can be captured in a matrix form as

$$\mathbf{z} = \mathbf{T}\mathbf{Z}. \tag{15}$$

where

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

is the *task matrix* defining the task realization in terms of the model end effectors.

We seek to define the mapping from articulators to end effectors and hence to tract variables. We first define a *layout matrix* used to relate pure articulators to end effectors:

$$\mathbf{Z} = \mathbf{L}\mathbf{y}. \tag{16}$$

In its generalized form, the layout matrix, $\mathbf{L}$, becomes

$$\mathbf{L} = \begin{pmatrix} -l^J_{TB} & 1 & 0 & 0 & 0 \\ -l^J_{TT} & l^{TB}_{TT} & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -l^J_{LL} & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that the elements on the superdiagonal are all equal to 1 as each of them simply accounts for the influence of the pure articulator associated directly with the given end effector. This leaves us with four tuneable parameters: $l^J_{TB}, l^J_{TT}, l^{TB}_{TT}, l^J_{LL}$. Three of the parameters associated with the jaw pure articulator – l^J_{TB}, l^J_{TT} and l^J_{LL} – depict the effect of the jaw movement on the tongue body, tongue tip and the lower lip end effectors, respectively. This effect of the position along the front-back axis of the end effector can be captured by setting the values of these transformation parameters in the following way

$$l^J_{TB} < 1 < l^J_{TT} < l^J_{LL}.$$

The remaining parameter l^{TB}_{TT} captures the influence of the tongue body movement on the position of the tongue tip. The effect of the anatomical constraint on the tongue tip movement can be formally approximated by setting

$$l^{TB}_{TT} < 1.$$

The end effectors themselves are related to the tract variables by Equation 15. We can combine Equations 15 and 16:

$$\mathbf{z} = \mathbf{TLy} = \mathbf{Ay},$$

which determines a single mapping from tract variables to pure articulators. We call this the *anatomy matrix*. In our simulations reported here, we use these values.

$$\mathbf{A} = \mathbf{TL} = \begin{pmatrix} -2/3 & 1 & 0 & 0 & 0 \\ -1 & 1/4 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 & -1 \end{pmatrix}, \tag{17}$$

Appendix 2: Model Dynamics

We define a time dependent *active anatomy matrix* $\mathbf{A}(t)$ and *active target vector* $\mathbf{z}_0(t)$ in the following way: for each gesture active at time t, the matrix $\mathbf{A}(t)$ contains the row of the model's anatomy matrix $\mathbf{A}$ associated with the tract variable connected with the given gesture. Similarly, $\mathbf{z}_0(t)$ will be a vector containing the gestural targets corresponding to the rows of matrix $\mathbf{A}(t)$.

If, for example, a gestural score at time 0.2 s prescribes a concurrent activation of gestures /a/ and /d/, the active anatomy matrix $\mathbf{A}(0.2)$ will contain the first and the second rows of the anatomy matrix $\mathbf{A}$ corresponding to the tongue body and tongue tip tract variables involved in the production of these gestures, i.e.

$$\mathbf{A}(0.2) = \begin{pmatrix} -2/3 & 1 & 0 & 0 & 0 \\ -1 & 1/4 & 1 & 0 & 0 \end{pmatrix}.$$

If $\mathbf{y} = (y_J, y_{TB}, y_{TT}, y_{LL}, y_{UL})^T$ is the vector of current pure articulator positions, then

$$\mathbf{z}(0.2) = \begin{pmatrix} z_{TB} \\ z_{TT} \end{pmatrix} = \mathbf{A}(0.2)\mathbf{y}$$

is the vector containing the values of the relevant active tract variables, and the corresponding active target vector $\mathbf{z}_0(0.2) = (z_{/a/}, z_{/d/})^T$ specifies the corresponding gestural targets.

We now describe the modification to Equation 10 required to appropriately link the stiffness coefficients of the tract variables to the inertial properties of the associated articulators. Here, repeated, is our Equation 10:

$$\ddot{\mathbf{y}} = \mathbf{A}^{*\mathbf{W}}\mathbf{M}_z^{-1}[-\mathbf{K}_z(\mathbf{Ay} - \mathbf{z}_0) - \mathbf{B}_z\mathbf{A}\dot{\mathbf{y}}]. \tag{10}$$

First, we expand both sides of Equation 10 by the (diagonal) pure articulator mass matrix:

$$\mathbf{M}_y\ddot{\mathbf{y}} = \mathbf{M}_y\mathbf{A}^{*\mathbf{W}}\mathbf{M}_z^{-1}[-\mathbf{K}_z(\mathbf{Ay} - \mathbf{z}_0) - \mathbf{B}_z\mathbf{A}\dot{\mathbf{y}}]. \tag{18}$$

The basic idea of our approach is to ensure that (i) the system's dynamical behavior arises directly from the distribution of mass within the set of articulators, and (ii) that the influence of the gestural stiffness coefficient $\mathbf{K}_z$ (Equation 6) is properly distributed among the pure articulators involved in the task realization.

The left hand side of Equation 18 correctly captures the distribution of forces to the individual pure articulators, given the articulator accelerations specified by the task-dynamics and the articulator mass matrix. Since the matrix $\mathbf{M}_y = \mathrm{diag}(m_1, \ldots, m_5)$ is diagonal, the force acting on the ith pure articulator is simply equal to $F_i = m_i\ddot{y}_i$.

In order to satisfy the second requirement, we must ensure that the quantitative evaluation of the pure-articulator stiffness coefficient

$$\mathbf{K}_y = \mathbf{M}_y\mathbf{A}^{*\mathbf{W}}\mathbf{M}_z^{-1}\mathbf{K}_z\mathbf{A} \tag{19}$$

of Equation 18, when recast to the tract variable space via the anatomy mapping $\mathbf{A}$, yields precisely the tract variable stiffness component $\mathbf{K}_z$ (Equation 6). This condition can formally be expressed as

$$\mathbf{A}\mathbf{K}_y\mathbf{A}^* = \mathbf{K}_z,$$

where $\mathbf{A}^*$ is a (right) pseudoinverse of the active anatomy matrix $\mathbf{A}$: $\mathbf{A}\mathbf{A}^* = \mathbf{I}$.

This is achieved if

$$\mathbf{K}_y = \mathbf{A}^*\mathbf{K}_z\mathbf{A}, \tag{20}$$

for *some* pseudoinverse $\mathbf{A}^*$ of the anatomy matrix $\mathbf{A}$, i.e., the matrices $\mathbf{K}_z$ and $\mathbf{K}_y$ are pseudo-similar.

If these matrices are related in this way, they have the same eigenvalues. An eigenvalue of a linear projection is a scalar which determines the ratio of scaling by this projection of a vector pointing in one of the principal directions determined by the projection. (These directions are expressed as eigenvectors of the given projection.) The relationship (20) thus guarantees that the *groups* of pure articulators involved in the realization of concurrently active gestures act proportionally to the gestural stiffness coefficients of the active gestures (the diagonal of matrix $\mathbf{K}_z$). At the same time this relationship ensures that the numerical values of gestural stiffness coefficients are interpretable as physical quantities scaled with respect to the physical properties of the vocal tract articulators.

If the task mass matrix $\mathbf{M}_z$ is set in the following way (note that at this stage this is the only component in the above equations that is "tunable", or not fully determined):

$$\mathbf{M}_z = \mathbf{A}\mathbf{M}_y\mathbf{A}^{*\mathbf{W}_1},$$

it can then be shown that

$$\mathbf{M}_y\mathbf{A}^{*\mathbf{W}_1}\mathbf{M}_z^{-1} = \mathbf{A}^{*\mathbf{W}_2}, \tag{21}$$

where $\mathbf{A}^{*\mathbf{W}_2}$ is the pseudo-inverse of the active anatomy mapping $\mathbf{A}$ computed using the weight matrix $\mathbf{W}_2 = \mathbf{W}_1\mathbf{M}_y^{-1}$. In the section on Pure Articulator Dynamics, we motivated the choice of the identity matrix as $\mathbf{W}_2$ (which fixes $\mathbf{W}_1$ as $\mathbf{M}_y$).

If this equation is substituted into Equation 19 it can be seen that the requirement expressed in Equation 20 is satisfied, and, moreover, the precise form of the pseudo-inverse matrix is identified.

Now, by substituting the identity (21) into Equation 18 we get the following instance of the pure-articulator dynamic system:

$$\mathbf{M}_y\ddot{\mathbf{y}} = \mathbf{A}^{*\mathbf{I}}[-\mathbf{K}_z(\mathbf{A}\mathbf{y} - \mathbf{z}_0) - \mathbf{B}_z\mathbf{A}\dot{\mathbf{y}}], \tag{22}$$

which we met before as Equation 12.

This dynamical system satisfies both requirements identified above. It imposes a genuine force-driven task dynamics on the system of model pure articulators. The mass matrix $\mathbf{M}_y$ is not obsolete, and the right hand side of each individual equation of the system contains a sum of task dependent forces acting on the appropriate pure articulator in order to achieve the given constellation of gestural targets in a manner proportional to gestural stiffness. Crucially, the evaluation of the left hand side of the equation results in the force vector driving the individual pure articulators with the physiologically motivated masses towards their collective targets in a manner determined by the appropriately scaled gestural stiffness parameters.

Extended by the oral cavity boundary element, the following form of Equation 22 defines the acceleration component of pure articulatory dynamics generated by the combined influence of all concurrently active tasks:

$$\ddot{\mathbf{y}}_{task} = \mathbf{M}_y^{-1}\mathbf{A}^{*\mathbf{I}}[-\mathbf{K}_z(\mathbf{A}\mathbf{y} - \mathbf{z}_0) - \left(\mathbf{B}_z + \frac{1}{10^9(|\mathbf{A}\mathbf{y} - \mathbf{z}_b|)^3}\right)\mathbf{A}\dot{\mathbf{y}}], \tag{23}$$

where $\mathbf{z}_b$ denotes the tract variable limits associated with physical boundaries of the vocal cavity, and cavity boundary constraints are modeled as increasing the degree of tract-variable damping. We use Equation 23 in our model to determine the task-oriented behavior of pure articulators.

This is the equation we use in our model to determine the task-oriented behavior of pure articulators.

This completes the description of the dynamics driving the performance of our model.

	Pure articulators				
	J	TB	TT	UL	LL
Masses (g)	55	250	50	30	30
Stiffness coefficients	1.5	1.5	1.5	1	1
S-R equilibrium positions	9	6	-5	1	-1

Table 5: Parameters associated with pure articulators.

		Gestural targets (mm)			
		/b/	/d/	/a/	/i/
Tract variables	z_{TB}			-9 (-11.6)	9 (11.6)
	z_{TT}		3 (5)		
	z_{LA}	0 (-4)			
Gesture stiffness coefs		15	17.3	5	5

Table 6: Parameters associated with tract variables. The task dependent equilibrium positions of the relevant tract variables, in brackets, lie beyond the actual physical positions of gestural targets used in evaluation of parsing cost component P.

Appendix 3: Parameters used in the simulations

Tables 5 and 6 list the numerical values of model parameters used in the simulations reported in this paper.

Table 5 contains the masses, stiffness coefficients and speech-ready state equilibrium positions of the system's pure articulators. The tongue body pure articulator, for example, acts on a mass of 250 g, it's speech-ready stiffness is 1.5 times the current value of system-wide stiffness k and it's speech-ready position is 6 mm above the position of the jaw. The masses listed in the table are on the diagonal of pure articulator mass matrix $\mathbf{M}_y$, the speech-ready equilibria form the vector $\mathbf{y_0}$.

Table 6 then lists the numerical parameters linking the tract variables with gestures defined in the system. In order to produce consonantal gesture /d/, the tongue tip tract variable, z_{TT} is attracted towards the gestural target 5 mm, the gestural stiffness is 17.3 times the current value of system-wide stiffness k. The position of the model alveolar ridge is set to 3 mm; the dynamical target of the tract variable, 5 mm, lies *beyond* the actual physical boundary.

The anatomy matrix $\mathbf{A}$ defined in Appendix 1 was used for mapping the pure articulator lengths to the positions of tract variables.

The gestural sequences reported in the paper are optimal for the following values of cost component weights: $\alpha_E = 1$, $\alpha_P = 4$, and $\alpha_D = 8$.

The reader can find the complete Matlab code used in this work, as well as the gestural scores used as starting points of the optimization process at ftp://cspeech.ucd.ie/pub/code/ETD.zip