Sequencing Embodied Gestures in Speech

Juraj Simko Fred Cummins [*†]

July 13, 2009

1 Abstract

The embodied character of cognitive motor systems has been reflected in recent models that in turn have greatly influenced understanding of their constitution and function. Embodiment has as a consequence that system behaviors must take appropriate account of energy expenditure and metabolic costs that are unavoidable in a physically realised medium. We here consider optimisation, presumed to result from both phylogenetic and ontogenetic processes, that can be used to constrain the space of potential degrees of freedom of a system, ensuring that the resulting action is efficient and smooth. To understand the emerging adaptations, it is necessary to factor in the properties of the physical and physiological substrate that anchor the system's goal-oriented performance.

[*]This work has been funded by Principal Investigator Grant number 04/IN3/I568 from Science Foundation Ireland to Dr. Fred Cummins, UCD, Dublin.

[†]Both authors are with the School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland, juraj.simko@ucd.ie, fred.cummins@ucd.ie

However, the embodied nature of speech production has been disregarded by most phonological research up to date. This leads to a failure in providing a coherent phonological grounding of a wide range of phenomena extensively documented by experimental phoneticians, in particular those associated with the relative timing of gestures (also called *gestural phasing*) in connected speech and its variability as found in different manners of speech. Existing phonological theories of sequencing rely on essentially external system-wide rules and principles or explicit dynamical constraints governing phasing to account for various suprasegmental properties and prosodic parameters of an utterance.

We introduce here a new and highly abstract modeling platform developed to investigate the embodied character of speech. The physically instantiated, second order dynamic nature of the system allows us to define and exactly evaluate various cost functions, which we hypothesise to play a role in efficient gestural sequencing. We investigate the general dynamical properties of the system, and identify a set of its high level, intentional parameters linked to the cost functions associated with its goal oriented performance. We show that the phenomena accompanying gestural sequencing, coarticulation, fluency and prosodic modulation, emerge as consequences of a non-trivially formulated efficiency constraint, thus providing a principled phonological account of phonetic reality.

2 Introduction

Speech production is a well rehearsed, sequential cognitive activity. Speaking brings about the precise coordination of multiple effectors belonging to a highly complex physical system stretching from diaphragm to lips. As with any form of skilled action, mastery of speech involves learning how to coordinate these diverse parts while respecting the constraints imposed by functional requirements, i.e. communication.

In his papers on Emergent Phonology (Lindblom, 1983; Lindblom, 1999), Lindblom put forward an idea that several general motor action and cognitive principles can, when mapped appropriately into the speech production and perception domain, shed a novel light on known phonological explanations and phonetic phenomena. Rather then being postulated as a system of external, representational laws, phonological phenomena *emerge* as consequences of these basic principles.

The basic constraint which shapes any cognitive embodied skilled motor activity is a requirement of *efficiency*. As we understand it, an *efficient* system is one that displays an energetically optimal trade-off between the conflicting demands of minimising effort in movement and maximizing perceptual clarity for the perceiver of speech. By adhering to this principle alone, the speech production cognitive system curtails the complexity of the task of generating and sequencing production primitives so that a continuous stream of speech sounds is produced. In other words, this principle helps to reduce dramatically the number of degrees

of freedom associated with a redundant motor action in general, and with speech production in particular.

In this view, the cognitive system is not seen as an autonomous disembodied controller acting *on* a physiological substrate and governed by abstract rules, but rather as containing the substrate, inseparable from it, acting *in accordance* with the physical constraints imposed by the environment, adapted to them, and taking full advantage of substrate's properties in assembling functionally determined coordinative structures for performing given tasks.

The usefulness of this approach for phonological research has been demonstrated for example in Dispersion-Focalization Theory (Schwartz et al., 1997), which provides an account of the distribution of individual vowels in a potentially continuous space of vocalic primitives. Schwartz and his collaborators re-interpreted stability criteria postulated in Stevens' Quantal Theory (Stevens, 1989) and perceptual distinctiveness criteria put forward in Lindblom's Dispersion Theory (Liljencrants and Lindblom, 1972) as complementary production and perception cost functions and used a simple optimisation to derive vowel distributions which closely matched various natural vowel systems. This result shows that the efficient global patterns pinpointed by the constrained optimisation have their real counterparts in existing phonologies. During the acquisition of their mother tongue the speakers can take advantage of the existence of these low-energy attractors in the production dynamics.

In our work we extend this approach to another dimension of speech production: that of the sequencing of gestures in time. We hypothesise that efficiency

requirements arguably influencing the distribution of speech primitives also play a crucial role in the way that these primitive actions are strung together in time when uttering a connected stream of speech.

Producing an utterance is to a large extent a *sequencing* task. It involves a precise phasing of the execution of primitive articulatory actions. The manner in which the actions are organised into patterns determines the content and quality of the acoustic output. The rules governing the ordering, the precise relative timing and overlaps of actions, and the high level parameters of their execution provide a descriptive framework for many types of phonological variation.

The phonetic manifestation of these sequencing variations encompasses the phenomena generically described as *coarticulation*. Coarticulation refers to the context-dependent manifestation of speech segments, or gestures, that arise as a direct result of the co-production of adjacent or nearby segments/gestures.

Prosodic variation influences coarticulation patterns in a wide variety of ways that have received much attention. In this work we shall focus primarily on the variations elicited by speaking rate manipulation.

Thomas Gay and his collaborators (Gay et al., 1974; Gay, 1981) noted that, in general, “the duration of segmental units, the displacement and velocity of articulatory movements, and the temporal ovelap between individual segments undergo nonlinear transformations during changes in speaking rate.” The nonlinear effects of rate changes were reported in different forms for both speech and non-speech motor actions. For example, they showed that consonants get shortened proportionately less then vowels in fast speech.

Nittrouer et al. (1988) and Nittrouer (1991) investigated the influence of rate changes on the phasing details of utterances. They showed that these changes result in changes of the *relative* phasing structure of an utterance. For utterances of /$C_1V_1C_2V_2C_3$/ form, the onset of the intervocalic consonantal gesture for C_2 – bilabial in (Nittrouer et al., 1988) and alveolar in (Nittrouer, 1991) – starts relatively earlier in the underlying vocalic cycle V_1–V_2 for fast speaking rate then for slow rate. In both cases the consonantal gesture also occupies a greater portion of the cycle at fast rates.

On the other hand, Cummins (1999) refined this general principle by observing that nonlinearities accompanying speaking rate changes are not manifested uniformly accross the duration of an utterance. In fact, the relative durational relationships between actions grouped within a suprasegmental unit (e.g. a syllable) remain more stable than those observed across unit boundaries.

Can these and similar variations be accounted for as natural consequences of adaptations of the cognitive system acting in accordance with the embodied neuromuscular system, our vocal tract, towards *efficient* production of the information-carrying stream of speech?

2.1 Background

To address this question, we present a developing modeling paradigm in which many of the superficial complexities associated with speech production are finessed, while many basic principles relevant to efficient sequencing and coordination in real time in a physically instantiated, embodied structure are

respected.

We draw inspiration for capturing the fundamental properties of an embodied articulatory system from three major spheres of research: the theory of optimality principles, motor action theory, and task dynamics. Phonologically, we ground our model of sequencing in the theory of Articulary Phonology and its Task Dynamic implementation.

Optimality principles play a vital role in the performance of skilled cognitive sensorimotor actions. Their appeal "lies in their ability to transform a parsimonious performance criterion into elaborate predictions regarding the behavior of a given system" (Todorov, 2004) (see also for an overview of the recent trends in this field). Both evolution and ontogenetic development craft behavioral systems under the constraint of efficient production, as costs associated with both producing and perceiving a message are always present. To account for gestural patterns resulting from the requirement of motor efficiency, we thus need to include a measure of energy expenditure associated with the system's performance in our model, i.e. we must have modeling access to the magnitudes of forces driving the movement of system constituents. We therefore have to build our model as a system embodied in the physical world, possessed of masses, and subject to physical constraints of continuity, impermeability, inertia, and the like.

Recent modeling attempts, e.g. (Anderson and Pandy, 2001), have explored a wide variety of individual cost functions relevant to different tasks. However, capturing the essential features of behavioral data typically requires the inclusion

of a number of distinct cost terms. We shall suggest a combination of three well-motivated terms in our cost function and show how the weights representing their prominence elicit the required lawful variation in speech production at multiple rates.

The low-cost patterns of speech production seem to play an important role in shaping many aspects of phonological structure. In the words of a leading proponent of this research paradigm, Björn Lindblom, phonological patterns can be seen as "products of cultural evolution adapted to universal biological constraints on listening, speaking and learning" (Lindblom, 2000). Our aim is to propose an interlinked system of such constraints balancing the production-oriented and listener-oriented influences. To be able to do that – in particular on the production-oriented side of the trade off equation – we must identify the right level of analysis allowing us to quantify the cost reflecting the elusive concept of articulatory ease.

Kinematic and dynamic properties of limb (and speech articulator) movement – such as velocity curves, changes in movement duration under different conditions of rate and extent – have been extensively studied, both experimentally (Ostry et al., 1987) and theoretically (Kelso, 1995; Ostry, 1986). Cooke (1997) suggested a second order dynamics (akin to a damped driven spring with mass) as a suitable approximation for modeling muscle-joint structures, and he showed that continuous change of its high level physical parameters, for example *stiffness*, lead to qualitative changes in the organisational form exhibited by the motor action systems.

This approach has been extended by the school inspired by the work of the great Russian physiologist Nicolai Bernstein which has developed the Equilibrium Point hypothesis of muscular action (Ostry and Feldman, 2003; Latash, 2008). The central idea behind the hypothesis is that muscular action is determined by shifts in the equilibrium position of the muscle dynamics. By equilibrium shift is meant the equivalent of a change in the resting length of a damped spring after which the load attached to the spring moves to a new position driven by forces dependent on the spring stiffness, damping and the load mass. Although the proponents of this approach explicitly stress the deficiencies of the simple mass-spring dynamics for modeling the highly non-linear characteristics of muscular action (Feldman and Latash, 2005), they nevertheless admit the methodological convenience of the global dynamic parameters like stiffness and damping in accounting for some broad patterns of the form of movement. Aware of the simplifications introduced by modeling low level muscular action this way, we presume that capturing the basic properties of speech articulators as essentially damped mass-springs with adjustable stiffness is sufficient for our main aim of demonstrating the role of efficiency principles for task sequencing in an embodied system.

A related methodology for describing the behaviour of complex motor systems has been introduced by Haken, Kelso and Bunz (1985) and by Kelso and Saltzman as Task Dynamics (Saltzman and Kelso, 1987). They introduced an abstract space of tasks performed by the system where the patterns of the task accomplishment (and not the underlying neuromuscular system) are captured by

the second order dynamics. The overall dynamics of the production system is thus determined functionally, by the behaviour towards the achievement of a given higher level goal, e.g. constriction of the vocal tract at a specific location. The task generates a synergy between the system's primitive components, groups them into a coordinative structure and, in effect, reduces the number of degrees of freedom of the unconstrained system. The dynamical patterns of the task and their interaction with a physically instantiated articulatory space account for many observed phenomena. This approach has been successfully adapted for speech production modeling.

Articulatory Phonology (AP) proposed by Catherine Browman and Louis Goldstein (Browman and Goldstein, 1991; Browman and Goldstein, 1992), postulates functionally defined, physically real dynamical events called *gestures* to be both fundamental units of information, i.e. phonological contrast, and primitive units of action, i.e. speech production. Every gesture imposes a set of goals (formation and release of a constriction at some place of the vocal tract) on the vocal tract in order to produce the desired phonological event.

Within AP, the behaviour of gestural primitives is captured in a top-down manner in three parallel levels of description: a *gestural score* level characterised by activation variables as functions of time, a vocal tract *task* level captured by abstract tract variables, and an *articulatory* (physically real, embodied) level described by model articulator variables. The gestural score represents organised patterns (constellations) of gestures participating in an utterance's production, and in particular the precise timing of their onsets and offsets relative to each

other – *gestural phasing*. Each tract variable represents the degree to which the goals associated with a gesture are achieved over time. Finally, an articulator variable shows the position of every articulator participating in the gestural target accomplishment. The degree to which a gesture has achieved its targets is captured by several (one or two) tract variables. Each tract variable is in turn associated, not necessarily exclusively, with several model articulators. For example, the TTCL (tongue tip constriction location) tract variable is linked to behaviour of tongue tip, tongue body and jaw, while the jaw at the same time affects the values of the LP (lip protrusion) tract variable. Therefore the mapping between elements of the articulatory layer (model articulator space) and those of the task layer (tract variable space) is of a many-to-many nature. This mapping is, however, seen as a relatively straightforward transformation between two coordinate systems.

In the standard implementation model of AP, Task Dynamics (TD) proposed by Saltzman and colleagues from Haskins Laboratory (Saltzman and Munhall, 1989; Saltzman, 1991), the dynamics of speech production is derived exclusively from the uncoupled dynamics of *tract variables* in the vocal tract task space. It is the motion of a gesture's tract variables, *not* the motion of the associated individual articulators, which is characterised dynamically. Given a gestural activation pattern and targets for each gesture, the precise manner in which these targets are achieved in time is the solution of a second order dynamical system where each tract variable is modeled as a mass spring with arbitrary (unit) mass. This solution is, as mentioned above, simply recast into the model articulator

coordinate system, thereby specifying the kinematics of the individual articulators participating in the target accomplishment. Crucially, the dynamics of the articulatory layer play no role whatsoever in determining the behaviour of this motor action system. Despite its dynamical nature and its emphasis on the importance of production constraints for a phonological theory, this approach models the speech production system in a top down manner. The physically real properties of articulators, e.g. their masses, are not represented in the TD implementation, which makes it impossible to track the force dependent quantities, such as energy expenditure, that are presumed to underwrite the efficient (energetically optimal) operation of the system.

The researchers participating in the AP project have managed to provide a coherent description of many phonetic and phonological phenomena. In particular, the theory successfully accounts for motor action robustness to external perturbations, naturally explains coarticulation as coproduction (the result of multiple gestures with total or partial activation overlap vying for control over articulatory system), or the existence of hidden gestures (which do not manifest themselves in the acoustic outcome of speech action, but are still present in the underlying production).

Without necessarily committing to all theoretical claims and implications of AP and TD, we will adopt their fundamental views and terminology: functionally defined gestures, the interacting levels of description and the inherently dynamical nature of speech production.

One area where we take a different stance to that of AP and its TD

implementation is in the sequencing and phasing of individual gestures, and in particular the observed dependency of the gestural score for an utterance on the manner of its production: that is, the dependency of the phasing relations among the gestures on the speaking rate and other sundry prosodic properties. Browman and Goldstein (2000) propose an essentially grammatical way for determining phase relationships between gestures participating in an utterance. Byrd et al. (2000) and Byrd and Saltzman (2003) proposed a more comprehensive theory allowing for flexibility in the phase relationships (windows of relative timing) and an additional layer of so called π gestures that capture some prosodic properties of an utterance.

Both of these approaches thus deal with the problem of determining the phasing of gestures by specifying essentially explicit descriptions and adding external rules governing phasing for various suprasegmental properties and prosodic parameters of an utterance. When seen as a cognitive system, the speech production apparatus is thus supposed to deal with the vast number of degrees of freedom associated with gestural phasing by a complex, representational structure without exploiting the physical and physiological environment in which it is embodied.

The main aim of our research is to identify constraints imposed on such an embodied articulatory system by its dynamics and efficiency requirements. That way we can investigate the general properties of the embodied articulatory space, describe its dynamics and propose a set of high level, intentional parameters which are at the disposal of a cognitive system and which elicit the qualitative

changes in organisational form characteristic of speech production at a range of rates.

In this paper we propose to exploit the inherent dynamics of the articulatory layer to investigate the space of gestural phasing. This space is delimited by physiology, physics and efficiency principles (arising from both phylogenetic evolution and ontogenetic development) and has to be mastered by and is at the disposal of a cognitive system participating in speech production. Speech conceived as an embodied motor action allows us to bring concepts like cost, production efficiency and optimality into the discourse on speech production. They then facilitate the identification of high level production parameters associated with the intentional control of speaking rate and production precision, which in turn help us to talk more precisely about elusive notions in phonetics and phonology like fluency, trade-offs between precision and rate, etc.

2.2 Speech production models

The human vocal tract is a very complex physiological system with over sixty muscles and several bone structures engaged in on-line shaping of its components, the *articulators*, to attain required aerodynamic shapes and movements. It consists of several more or less independent and loosely defined articulatory subsystems: tongue and palate, jaw and lips, velum, vocal folds, to name a few. Within each of these subsystems, articulators are relatively strongly linked by the anatomical organization of the tract.

To produce a velar stop, for example, the back part of the tongue body (dorsum)

must achieve a contact with the rear portion of palate (soft palate). This movement obviously impacts the behaviour of all articulators grouped within the tongue and palate subsystem: it strongly limits the possible shapes of the entire tongue body and its positions relative to the palate and to a lesser extent also the absolute position of the (anatomically quite flexible) tongue tip and its position relative to, e.g., the alveolar ridge or teeth.

On the other hand, articulators from distinct subsystems, e.g. tongue body and vocal folds, can act quite independently: a velar stop can be produced with vocal folds vibrating (producing a voiced consonant, e.g. /g/) or with vocal folds relaxed (producing a voiceless /k/). This flexibility is exploited by the speech production system; the relative functional independence of subsystems allows for a combination of their activity patterns resulting in various phonologically contrasting speech sounds. With some caution it might be said that during speech production the articulators within each subsystem are coupled strongly by anatomy and relatively weakly functionally, while articulators from different subsystems are strongly functionally coupled and comparatively weakly linked by anatomy.

There are several physiologically motivated, articulatory models of the vocal tract (Maeda, 1982; Boersma, 1998; Iskarous et al., 2003). These models focus on the very complex task of articulatory speech synthesis, i.e. the accurate translation of changing shapes and movements of the vocal tract into acoustic space. The input to these models is a sequence of parameters defining the shapes and positions of model articulators, which are then used to synthesise the acoustic output. The

kinematics of the articulators are *fully* determined by the input sequence. The models thus abstract away from the dynamical properties of their components, the masses and forces acting on the model articulators – the vocal tract they deal with is thus disembodied with respect to its dynamical properties.

These models are then traditionally used in modeling higher level characteristics of speech production and perception (Guenther, 1995; Howard and Huckvale, 2005). As mentioned in the case of AP above, although the connected theories are successful in accounting for many matters related to speech production, acquisition and even in a projection of speech related processes into human brain operation, they inevitably stop short of explaining phenomena linked to the embodied nature of speech production, e.g. articulatory efficiency, high level parametrisation of speaking rate and articulatory precision control, all associated with the emergent lawfulness of gestural phasing.

In recent years, several detailed models of the entire vocal tract or its subsystems taking the embodiment of vocal tract seriously have been proposed (Perrier et al., 2000). These models aim to get hold of the enormous complexity of physics and physiology behind the real vocal tract and its ability to produce speech. As a consequence, the synthesis of an acoustic output or detailed stream of vocal tract configurations for even a short utterance takes very long time even on the fastest computers available to date. Much as they are useful for testing the predictions of current production theories, these models can not be realistically employed for our task of describing the properties of the entire physically relevant *space* underlying speech production. To do that we need to be able to find optimal,

efficient gestural constellations which in practise means exploration of a search space by running thousands of slightly varying productions of each utterance.

3 Abstract Model

Our approach is thus to build a physiologically and physically motivated model of the speech production system, or, more precisely, of its supra-glottal subsystem in the above sense. Instead of trying to capture the superficial details of human vocal tract anatomy and the acoustic principles behind soundwave generation, we focus exclusively on the essential dynamical properties of an embodied motor action system incorporating only high level features of the human vocal tract that constrain speech production. The primitive constituents of the model must be relevant for our aim to capture adaptations that result in efficient, low energy sequencing of motor actions.

Therefore, we decided, first, to leave out the synthesis of acoustic output. Our model generates purely articulatory output – kinematic traces of the model articulators. Although some important perception phenomena are related to nonlinearities of the articulatory-to-acoustic mapping, we presume that the perceptual quality of the system's output is directly related to the articular precision with which it approaches a given target. Second, we leave out the details of anatomical aspects of vocal tract. The articulators of our model are, for now, purely abstract. They do not map directly onto specific articulators (tongue tip, lips...), but instead they capture high-level dynamical properties we are

interested in. For a given modeling task we can adjust the details of our model to reflect particular properties of and relations between speech articulators relevant for the given circumstances.

At some level of abstraction, the vocal tract can be seen as a (particularly complex) system driven by intentionally imposed muscular force impulses. As suggested by research in motor action, such systems can in principle be modeled using an appropriate (presumably, again, very complex) set of second order differential equations. Our approach is to turn this premise the other way around: take a relatively simple system with second order dynamics (yoked pendula) and investigate to what extent its behaviour can shed light on the constraints underlying human speech production.

As we aim to account for phenomena related to gestural sequencing and phasing, one of the high level features of speech production we need to capture by our modeling approach is a different nature of production of vowels and consonants. As experimentally first documented by Öhman (1966) and subsequently conceptualised by, e.g., Fowler (1983), vowels and consonants fall into two distinct natural articulatory classes. Vowels, in contrast to consonants, are produced by relatively slow, configurational movements of articulators. The vocalic gestures have broader targets and they engage to a large extent only three supraglottal articulators: tongue dorsum, jaw and lips. The consonantal gestures (or at least stop consonants), on the other hand, are generated by rapid, ballistic movements of articulators and involve their mutual collisions forming appropriate full constrictions of the vocal tract. These differences often lead to

postulating two separate articulatory tiers; during production, the consonatal tier is superimposed on the vocalic one. In Section 4, we will show how we can account for the facts underlying this hypothesis and also what properties of the system are related to the emergence of the separate articulatory tiers.

Rather then offering a full, exhaustive model of the speech apparatus, we aim to provide a modeling paradigm which allows us to build a succession of simple setups capturing various tangible phonological and phonetic phenomena. As described below, the approach allows for expanding and refining the basic setup to accommodate diverse hypotheses, test them and to serve as an intuition pump for thinking about sequencing and fluency.

The articulators are represented by pendula driven by torsion springs. Their kinematics is given by the solution of the following non-linear differential equation:

$$m\ddot{\theta} + b\dot{\theta} + k(\theta - \theta_0) = 0, \tag{1}$$

where m is the moment of inertia of the pendulum, k is the torsion spring coefficient, $b = 2\sqrt{mk}$ is the critical rotational damping parameter, and θ_0 is the resting angular deflection of the pendulum. The dynamics prescribed by Equation 1 is equivalent to the dynamics of damped mass-spring behaviour; the reason for choosing the torsion spring driven pendula instead of springs is to enclose their action in a compact space by warping a sequence of springs around in a circle. Because the effect of this decision can be seen as a purely geometrical transformation, for the sake of compatibility with the tradition of approximating

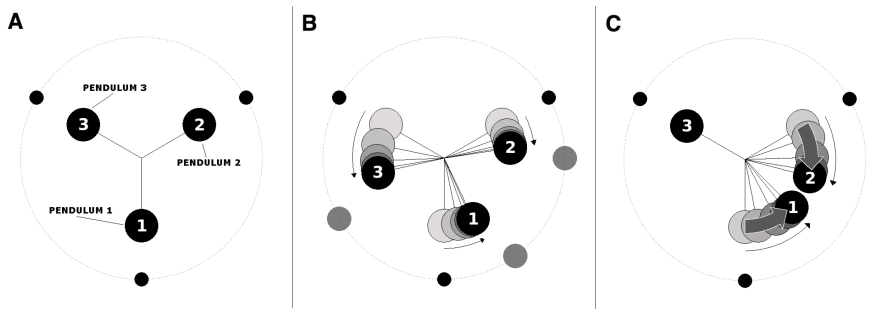

Figure 1: Basic setup (A) and a production of vocalic (B) and consonantal (C) gestures.

the neuro-muscular action with damped springs we shall use the spring equivalents when referring to the coefficients of Equation 1, i.e. we shall call m, b and k mass, damping and stiffness, respectively. Similarly, we shall use the term force when taking about the torque driving the pendulum action.

The basic setup of our abstract model is shown in a rest state in Fig. 1A. It consists of three torsion spring driven pendula hung on massless rods from a common point. Each pendulum has its own mass m_i, stiffness k_i, angular equilibrium position $\theta_{0i}, i = 1, 2, 3$ (in the above sense). The pendula are critically damped through the damping coefficients b_i to avoid overshoot and oscillation (the critical damping coefficient is calculated during an utterance production to dampen the forces elicited by all active gestures). Each pair of bobs exhibits mutual repulsion force R_{ij} when the bobs approach one another. The force R_{ij} is negligible for all but very small angular distances, and infinite for zero distance.

For various setups, it is possible to impose further second order dynamical

constraints on the system. Two pendula can, for example, be joined (linearly coupled) by a spring of a given length and stiffness to simulate a partial anatomical link between the two abstract articulators (for example, as in the case of tongue tip and tongue dorsum in the human vocal tract).

The pendula are the abstract articulators of our vocal tract model, presumably belonging to a single anatomically constrained subsystem. They can be acted upon by both configural (vocalic) and ballistic (consonantal) gestures and there exists a one-to-many relation between articulatory goals and articulators. Their movements are subject to physical constraints, and we can specify the dynamics governing their movement in a relatively simple fashion.

The system's rest dynamics describes a *speech ready* state of the articulatory system. Rather then representing an idle, purely anatomical resting positions of the articulators induced by, e.g., gravity (for example, the tongue lying flat against the jaw, the lips closed, teeth clinched together), our speech-ready equilibrium position is the state in which the articulatory system is "receptive", pre-configured for action elicited by speech gestures. This state is presumably established during a speaker's native vowel space acquisition and fine-tuning and is the state from which the vowel targets can be reached most economically (cf. (Barry, 1998)). Phonetically, the resting equilibrium corresponds to the configuration of the vocal tract producing schwa /ə/.

Vocalic gestures (syllabic nuclei) are implemented within the system as sets of additional absolute equilibrium positions involving one or more pendula.

In order to produce a "vowel" defined as an absolute articulatory configuration,

extra centres of attraction force induced in the same way as the resting equilibria can be switched on around the system at angular positions different to those for the resting position, each acting on one pendulum. We call these attractor tuples (one attractor position for each pendulum) the *attractor sets*. The magnitude of the attractive forces is larger then that of the resting attractor forces, so that pendula get displaced from the resting position towards the vowel configuration as illustrated in Fig. 1B. This is achieved by setting a proportionally higher value of the stiffness coefficient than the value for the resting attractor i.e. by multiplying the resting stiffness of each engaged articulator by a vocalic gesture gain. If an attractor set is switched on, after some time, and with a velocity proportional to the stiffness and mass of each pendulum involved, the engaged pendula move to stable positions, with their angular deflections close to those of the attractor set. (Not exactly the same, as the default resting position attractor set is still acting on them, although with a relatively weaker force.) When the attractor set is then switched off, the pendula slowly return to their respective resting positions.

Consonantal gestures are represented in a distinctly different fashion. Each is initiated by mutual attraction forces – forming a *consonantal attractor set* – between a pair of pendula, proportional to their distance and depending on each model articulator stiffness. The pendula thus move towards each other (Fig. 1C), until their mutual repelling forces equal the driving forces. This collision of two pendula constitutes an articulatory target ("closure"). The system remains in this closure state until the attractor set is disactivated. Again the path taken to reach

closure and move away from it, and the path taken during the closure duration, will depend on past state and future goals.

3.1 Formal definition

In this section we shall present a formal, mathematical description of our model. The resting dynamics of the system can be expressed by a system of three differential equations. The vector $\ddot{\boldsymbol{\theta}}_\ominus$ describing the acceleration of the model articulators is given by the equation

$$\ddot{\boldsymbol{\theta}}_\ominus = \mathbf{M}^{-1}\left(-\mathbf{B}\dot{\boldsymbol{\theta}} - k\mathbf{K}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \mathbf{R}(\boldsymbol{\theta}) - \sum_z k\kappa_z \mathbf{C}_z^*(\mathbf{C}_z\boldsymbol{\theta} - \theta_{z0})\right), \quad (2)$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ is a vector of the three articulator positions, $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \theta_{30})^T$ the vector of their resting equilibria, $\mathbf{M} = \mathrm{diag}(m_1, m_2, m_3)$ is the diagonal mass matrix with pendula bob masses on its diagonal, $\mathbf{B} = \mathrm{diag}(b_1, b_2, b_3)$ is the diagonal damping matrix with the damping coefficients on its diagonal (as mentioned above, this element is calculated to dampen all forces active at time, not only the resting one), and $\mathbf{R}(\boldsymbol{\theta}) = (r_{ij}(\boldsymbol{\theta}))$ is the repulsive force matrix, each $r_{ij}(\boldsymbol{\theta})$ being the magnitude of the repulsive force R_{ij} for the angular deflections $\boldsymbol{\theta}$, $r_{ii}(\boldsymbol{\theta}) = 0$, for all $i, j = 1, 2, 3$.
Each optional z in the leftmost sum element imposes a possible linear coupling (spring) between two articulators, with stiffness $k_z = \kappa_z k$ (generally much smaller then any k_i) and length θ_{z0}. If z is a coupling between ith and jth articulators $i < j$, the 1×3 matrix $\mathbf{C}_z = (c_{z1}\ c_{z2}\ c_{z3})$ is defined as follows:

$c_{zj} = 1, c_{zi} = -1$ and the third element is set to 0. This matrix represents a projection from the space of our three articulators into the simple task space containing the only task of keeping the $\mathbf{C}_z\boldsymbol{\theta} = \theta_j - \theta_i$ constant (equal to θ_{z0}). The 3×1 matrix $\mathbf{C}_z^*$ is the Moore-Penrose pseudoinverse of $\mathbf{C}_z$ mapping the status of the task achievement back to the articulatory space. This approach is analogous to the mapping between task space and articulatory layer used by the Task Dynamic implementation of Articulatory Phonology (Saltzman and Munhall, 1989).

The expressions $k\mathbf{K}_0$ and $k\kappa_z$ capture an important design decision incorporated in our model. To reduce the number of the parameters of the optimisation process, the various system's stiffness parameters (the rest dynamics stiffness k_i of the ith articulator, each coupling spring stiffness k_z, and gestural stiffness described below) are all related to each other in a linear fashion. In other words, each $k_i = \kappa_i k$ and each $k_z = \kappa_z k$, where k is the overall system-wide stiffness; κ_is and κ_zs are stiffness *coefficients* of the pendula and coupling springs, respectively. The matrix $\mathbf{K}_0 = \mathrm{diag}(\kappa_1, \kappa_2, \kappa_3)$ is a matrix containing the rest dynamics stiffness coefficients on its diagonal. These coefficients are parameters of a given setup and remain constant for the given modeling task. It is possible, however, to adjust the overall system stiffness k and thus control how swiftly the pendula react to the application of the resting forces and forces induced by gestural activations.

An active vocalic attractor set v imposes an additional acceleration on articulators

$$\ddot{\boldsymbol{\theta}}_v = -\mathbf{E}_v \mathbf{M}^{-1} \kappa_{\text{voc}} k \mathbf{K}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_v), \tag{3}$$

where $\mathbf{E}_v = \text{diag}(e_1, e_2, e_3)$ is a diagonal matrix prescribing which articulators are engaged in the production of the vowel v (e_i equals 1, if the ith attractor is involved, and 0 otherwise), and $\boldsymbol{\theta}_v$ is the vector of the vocalic attractor equilibrium positions, and κ_{voc} is the vocalic stiffness gain.

Each attractor set represents a context-free vowel target, but the approach into and path from the configuration will be context sensitive, depending on past states of the system and future articulatory goals.

The consonantal collision gestures introduce a target-driven linear coupling between pendula, equivalent to the coupling elicited by the optional anatomically inspired springs discussed in the previous section. Thus, the acceleration imposed by a consonantal gesture c acting on ith and jth pendula ($i < j$) can be formally expressed as

$$\ddot{\boldsymbol{\theta}}_c = -\mathbf{M}^{-1} \kappa_{\text{con}} k \mathbf{K}_0 \mathbf{C}_c^* \mathbf{C}_c \boldsymbol{\theta}. \tag{4}$$

The task-articulator mapping matrix $\mathbf{C}_c$ is defined as the projection matrix $\mathbf{C}_c$ in the previous section, i.e. $\mathbf{C}_c = (c_{c1}\ c_{c2}\ c_{c3})$ where $c_{cj} = 1, c_{ci} = -1$ and the third element is set to 0, $\mathbf{C}_c^*$ is its Moore-Penrose pseudoinverse, and k_{con} (set to 4 for the models presented in this paper) is the relative stiffness coefficient for consonantal gestures. The equilibrium distance for the consonantal task is 0.

For example, for the consonantal gesture c_{13} between pendula 1 and 3, $\mathbf{C}_{c_{13}} = (-1\ 0\ 1)$. The pseudoinverse $\mathbf{C}^*_{c_{13}} = (-\frac{1}{2}\ 0\ \frac{1}{2})^T$ and the product

$$\mathbf{C}_{c_{13}}\mathbf{C}^*_{c_{13}} = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

The gestural dynamics imposed on the pendula 1 and 3 is then

$$m_1\ddot{\theta}_{c1} = -\frac{1}{2}\kappa_{\text{con}}\kappa_1 k(\theta_1 - \theta_3) \tag{5}$$

$$m_3\ddot{\theta}_{c3} = -\frac{1}{2}\kappa_{\text{con}}\kappa_3 k(\theta_3 - \theta_1). \tag{6}$$

The pendulum 2 is not influenced by the consonantal gesture is this case. Unlike the vocalic gestures defined by Equation 3, the consonantal gestures introduce a coupling between the articulators. This task oriented coupling (as opposed to the coupling reflecting anatomical constraints introduced in Equation 2) exemplifies an important difference in our approach to modeling vowels and consonants. While vowels are seen as *absolute* positional configurations of model articulators, the consonantal gestures impose a mutual coordination between pair of articulators acting in synergy.

3.2 Activation functions

In principle, each gesture can be triggered independently of any other and multiple gestural activation patterns can partially or totally overlap. Because production targets are defined in different ways for consonants and vowels, co-production of gestures is possible.

Each (vocalic or consonantal) gesture defined for the model can be switched on and off in during model's execution. These gestural activation patterns are modelled via activation time functions. For a given gesture g, the value of the activation function $a_g(t)$ at time t is set to 1 if the gesture g is active at time t, and to 0 if it is not active. The step-wise shift from 0 to 1 in the activation function a_g marks the onset of the gesture's g production, the step-wise shift form 1 to 0, its offset. The ensemble of all activation functions (one per each gesture defined for the model) is equivalent to the gestural score of Articulatory Phonology. The activation patterns are illustrated in the top part of Fig. 3.

The overall behaviour of our model is thus given by the equation

$$\ddot{\boldsymbol{\theta}} = \ddot{\boldsymbol{\theta}}_{\text{ə}} + \sum_{v} a_v(t)\ddot{\boldsymbol{\theta}}_v + \sum_{c} a_c(t)\ddot{\boldsymbol{\theta}}_c \tag{7}$$

incorporating Equations 2,3 and 4. The two summation expressions range over all vocalic and consonantal gestures defined for the model, respectively. The rest attractors (speech ready) are always on, while the gestures influence the system's dynamics according to their activation functions.

Our model is in many respects similar to the Task Dynamic implementation of

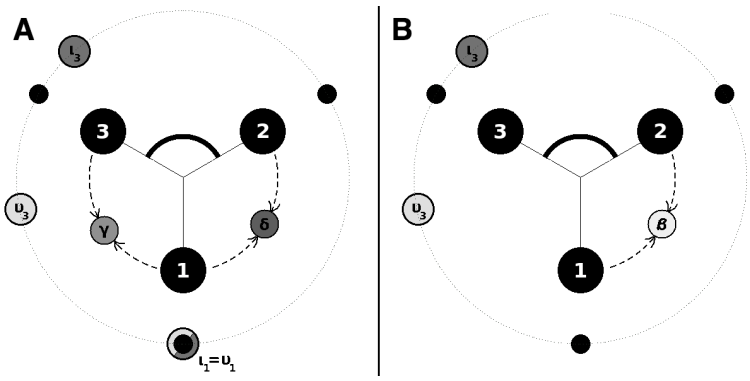

Figure 2: "Linguo-palatal" (A) and "linguo-labial" (B) setups of our model. The filled circles on the perimeter represent target equilibrium positions for given vowels and given pendula, e.g. υ_3 represent the υ target for pendulum 3. The filled circles placed between pendulum rods represent consonantal targets. For both setups, the masses of pendula 1,2 and 3 are 100g, 40g and 50g, respectively, their stiffness coefficients are 3, 1.8 and 2.5, respectively. The resting length of the spring joining pendula 2 and 3 is 140 degrees, its stiffness coefficient is 0.2. The definitions of vocalic and consonantal attractors are given in the text.

Articulatory Phonology described in Section 2.1. The important difference, however, is that the dynamics in the presented model originates in the embodied articulatory layer, i.e. the model's behaviour reflects the physical properties of its constituents (masses, stiffness coefficients, etc.). As we have already argued, this aspect is vital for our main intention of modeling the influence of efficiency requirements on speech dynamics in general and on the gestural phasing in particular.

3.3 Model parameters, inputs and outputs

Fig. 2 shows a schematic representation of possible model setups. They map, tentatively, to a linguo-palatal and linguo-labial articulatory subsystems of the human vocal tract, respectively. In the "linguo-palatal" setup (fig. 2A), the pendula 2 and 3 play the role of a tongue tip and tongue dorsum respectively, they are linked by a spring with relatively low stiffness, and pendulum 1 represents the top of mouth, its left hand side the velum and its right hand side the alveolar ridge. There are two vocalic attractors sets, each involving two pendula: the set labeled ι with target angular deflections of 0 and 220 degrees of pendula 1 and 3 respectively, and one labeled υ with target angular deflections of 0 and 280 degrees of the same two pendula. The two consonantal gestural targets are labeled δ (collision of "tongue tip" pendulum 2 with "alveolar" side of pendulum 1) and γ (collision of "tongue dorsum" pendulum 3 with the "velum" side of pendulum 1).

For the sake of comparability of simulation results, both presented setups share their quantitative characteristics, and differ only in the definition of their vocalic attractors. The 'linguo-labial" setup (fig. 2B) exhibits less constrained relationship between the vocalic and consonantal gestures. The pendula 1 and 2 represent lower and upper lip, respectively, and pendulum 3 simulates the tongue dorsum. The two vowels defined for this setup are determined by position of pendulum 3 only. The only consonant, labeled β (collision of the two labial pendula) models a bilabial stop.

The masses of pendulum bobs (matrix $\mathbf{M_0}$), the stiffness coefficients ($\mathbf{K_0}$, κ_zs,

κ_vs and κ_cs), the equilibrium vectors $\theta_\mathbf{0}$ and $\boldsymbol{\theta}_v$s, the coupling spring lengths θ_zs and vocalic gesture engagement matrices $\mathbf{E}_v$s are all *parameters* of a given model setup. They represent the physiological properties of the vocal tract, as well as the articulatory properties of a given phonetic space. They remain constant not only during each single simulation of an utterance production, but also during our entire exploration of the phasing behaviour of the model.

The remaining two parameters of the model act as independent inputs to the model. The first one is the collection of activation functions $a_g(t)$ for each gesture g defined in the model, i.e. a multidimensional input stream containing the activation patterns of predefined vocalic and consonantal attractor sets. As we have already mention, it is also possible to modulate the stiffness parameters of all constituents of the model via overall system stiffness k. (At the current stage of the model's development the overall stiffness is kept constant during an utterance production.) These inputs and their components have to be coordinated to elicit speech-like organization and kinematics.

The solution of the differential equations describing the model setup parameterised by the inputs is obtained by numerical approximation (implemented in Matlab) and yields angular positions of pendula in time (see Fig. 3) representing the articulatory behaviour .

These functions are then used to calculate the degree to which the vocalic or consonantal phonetic segments (targets of the gestures) are achieved at any given moment. As mentioned above, the output of the system described by Equation 7 is restricted to the traces of pendulum bob positions in time; there is no actual

sound production. For vowels, the distance of the present state from a given attractor set, absolute or relative, is inversely proportional to the *prominence* of targeted vowel. Similarly, the prominence of a consonant is inversely proportional to the mutual distance of two pendula engaged in the consonant's production. Prominence ranges from 0 to 1, with 1 representing perfect achievement of the target configuration. The prominence trace $p_g(t)$ of the gesture g in time plotted for given inputs can be interpreted as a vocal tract variable as employed by AP and TD. As mentioned above, the state of these "tract variables" capturing gestural target achievement is in our model derived from the dynamics of the articulatory layer of description and not the other way round, as in TD. The prominence $p_g(t)$ does not express the degree of achievement of the task prescribed by an independent task dynamics, but rather it captures the degree of realisation of the given segment *derived* from the overall behaviour of the system and its embodied dynamics parametrised by activation patterns and overall system stiffness. Unlike TDs tract variables, prominence traces are not solutions of an uncoupled second-order differential system. After evaluating all prominence functions for defined gestures, the phonetic segment with the highest prominence is deemed to be produced at any given instant.

3.4 Cost evaluation

The relative simplicity of our abstract model's design allows us to track details of various physical quantities involved in its speech-like activity: duration of

gestures, groups of gestures and entire utterances, kinematic characteristics (velocities, acceleration) of the movement of model articulators, and, importantly, the forces eliciting the dynamics of the system. These and similar measures are closely linked to the natural notion of *cost* associated with motor action, which, in turn, plays a crucial role when talking about efficiency, fluency and, as argued below, intentional control of high level behaviour of the system. Currently, we are using our model to evaluate three types of cost naturally related to speech production: *force expenditure cost* E – the overall physical effort involved in utterance production, utterance *duration cost* D influencing speaking rate and phonological relevance (salience) of the utterance's suprasegmental entities; and *parsing cost* P linked to the receiver's effort needed to parse the produced utterance. These three cost components allow us to quantitatively model the trade off between articulatory ease (represented here by E and D) and perceptual clarity (P). As we shall show below, all these expenditures are defined as functions of the inputs of our model, gestural activation functions and overall system stiffness.

Force expenditure. The force expenditure cost E is related to the concept of articulatory ease. It has, by far, the most elaborate definition of the three cost components and will be described first. Being a naturally occurring dynamical motor action driven by muscular activity, speech production is shaped by the requirement to minimise, at least in long term, the overall force used by its physiological components during speech production. This force needs to be computed.

The best approximation available for the overall force actively driving an utterance production is the value of what we call the *absolute gestural impulse* J: the integral over the duration of the utterance of the sum of magnitudes of all forces that act on model articulators with the exception of repulsive forces, which serve merely to simulate the solidity of pendula, and forces elicited by purely anatomical coupling between the articulators. That is, for production starting in time 0 and completing in time T

$$\begin{aligned}\mathbf{J} = \int_0^T (|\mathbf{B}\dot{\boldsymbol{\theta}}| + |k\mathbf{K}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0)| + \sum_v a_v(t)\,|\mathbf{E}_v\kappa_{\text{voc}}k\mathbf{K}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_v)| + \\ + \sum_c a_c(t)\,|\kappa_{\text{con}}k\mathbf{K}_0\mathbf{C}_c^*\mathbf{C}_c\boldsymbol{\theta}|)dt\end{aligned} \tag{8}$$

is the vector $(J_1, J_2, J_3)^T$ of force expenditures per articulator. The overall absolute gestural impulse is then

$$J = J_1 + J_2 + J_3. \tag{9}$$

This way of evaluating the expenditure of force in a motor action system reflects our outlook that behind *every* active force (positively or negatively) influencing the dynamics of the production system there is an active muscular structure. For example, even though most of the time the always-on speech ready dynamics acts against the forces elicited by active gestures, we presume that there is an anatomic structure (possibly one of the pair of agonist/antagonist muscles, or even a subsystem of their fibres) engaged in the pull towards the resting

equilibrium. Also, we presume that there is a similar muscular structure engaged in generating the critical damping forces. During the development of our model we experimented with various definitions of the overall force expenditure. The hypothesis presented here is supported by the insight we gained during this phase.

Further, we presume that the system increases its overall stiffness k by increasing the tension (tone) of muscles acting on each articulator in opposite directions with magnitude proportional to k. We approximate the force required to maintain the required system's stiffness by the stiffness value k itself. Thus, in our model the force expenditure cost is defined simply as

$$E = J + k. \tag{10}$$

Being a measure of the articulatory effort this cost function has presumably an universal influence on the speech production. The remaining two cost concepts are more directly related to our aim of finding the efficiency requirements responsible for phenomena associated with speaking rate adjustments and related sequencing variations.

Parsing cost. One of the strategies people adopt when they speak fast (while maintaining or even reducing the effort, the force expended during speech production) is target undershoot (Lindblom, 1963). The speakers compromise the precision with which some or all of the required gestural targets are achieved, in particular those associated with vocalic segments; they hypoarticulate.

Presumably, however, this target undershoot increases the *receiving* party's cost associated with parsing the utterance.

Speech is a social activity, so its production and perception aspects are inseparable – mere articulation is only a part, albeit important, of an utterance production; the utterance must be successfully parsed by the receiving party to be fully realised. More undershoot generally implies higher parsing cost. Therefore, we include a measure of the receiver's parsing cost in the overall cost of utterance production. The parsing cost P is thus defined as a sum of undershoots over all relevant phonetic segments produced in an utterance, i.e. distances of their peak prominences from the ideal, perfect prominence value of 1.

The requirement of minimising the force expenditure cost and the parsing cost pose conflicting constraints on the speech production: lowering the force expenditure increases the cost of the parsing the resulting utterance and vice versa. Efficient production of a given utterance is thus the result of a weighted compromise between these conflicting demands. This observation and its connection to phonetic variation is the basis of Lindblom's Hyper-Hypoarticulation Theory (Lindblom, 1990). Here we provide a quantitative expression of his predominantly theoretical outlook.

Duration cost. When choosing the speaking rate and other prosodic aspects of speech, overall utterance duration, or the duration of some well defined functional suprasegmental units (e.g. syllables) seems to be a natural parameter to control. The first approximation of the duration cost value D is the overall duration T of the utterance.

The same sequence of phonetic segments can be often uttered in two distinct ways, consider for example two utterances /di—did/ and /did—id/, where the dash marks a syllabic boundary. The requirement of salience of the differently organised syllables in these two utterances poses stricter temporal constraints on the inter-gestural phasing of the segments lying within the same syllable then on those lying across the syllabic unit boundary. The phasing patterns governing the production of a syllable are more stable then inter-syllabic ones (Byrd, 1996). We hypothesise that this suprasegmental salience constraint emerges as a consequence of the duration cost function being unevenly distributed over utterance production. To test this hypothesis, we designed the duration cost function D to reflect the required suprasegmental (e.g. syllabic) structure of an utterance.

The first modeling approximation of duration cost function conceived this way is an integral

$$D = \int_0^T \pi(t)dt \tag{11}$$

where the step function π is defined in the following way: $\pi(t)$ is set to 1 if t falls between onset and offset of a suprasegmental unit, e.g. a syllable, and to a constant value π_c, $0 \leq \pi_c \leq 1$, otherwise[1]. Thus the duration of periods when π has less then its maximal value 1 "counts less" then the duration of the intra-segmental periods. In the simulations presented in this paper the onset and offset of the CV syllables is associated with the moment of maximal prominence

[1]This approach to imposing a suprasegmental structure upon an utterance is inspired by Byrd's and Saltzman's prosodic π-gestures mentioned above.

of C and the moment of maximal prominence of V, respectively, and *not* the onsets and offsets of the underlying gestures. Thus the function π reflects the surface structure of the utterance, and therefore influences the gestural phasing indirectly.

All cost components E, D and P defined above are functions of a given utterance (sequence of gestures) in general, and of the manner of its production in particular, i.e. of the precise gestural constellation details, plus the overall system's stiffness related to system's agility to attain the required gestural targets. We do not claim that the three cost measures introduced here provide an exhaustive list of constraints behind efficient speech production. On the contrary, we are fully aware there are many other possible candidates for cost functions which can be used to account for phonetic and phonological phenomena, e.g. jerk (maximal acceleration of model articulators) or work, to name a just a few. As argued in the rest of this paper, our selection nevertheless seems to be the right one to shed novel light on the aspect of speech production under scrutiny here, i.e. fluent, efficient gestural sequencing and its dependency on intentionally adjusted speaking rate and articulatory precision.

3.5 Optimisation

The central hypothesis behind our work is that the requirement of optimal behaviour, i.e., the drive towards the minimisation of all three cost measures E, P and D, reveals general properties of the space of efficient, natural gestural constellations.

Having a vector of three cost measures to be minimised simultaneously, we are presented with a multiobjective optimisation problem. We approach it in a standard way – a weighted sum strategy – and convert it into a scalar optimisation problem by considering a weighted sum of all objectives:

$$C_{\boldsymbol{\alpha}} = \alpha_E E + \alpha_P P + \alpha_D D,$$

where $\boldsymbol{\alpha} = (\alpha_E, \alpha_P, \alpha_D)$ is a vector weighting the cost components, and hence in- stantiating a specific trade-off between production and perceptual costs.
The small changes in gestural constellation and overall stiffness value which lower the value of one of the cost functions under consideration generally cause an increase of the value of one or both remaining cost measures. For example, a smaller undershoot (decrease of the parsing cost P) can be achieved by increasing the overall stiffness of the system (increasing the force expenditure cost E) and/or by lengthening the duration of gestural activation (in effect increasing the durational cost D). This important property of the selected system of cost functions guarantees that, for a given weight distribution $\boldsymbol{\alpha}$, there exists a "compromise" solution – a gestural constellation and overall stiffness value – of the given optimisation problem of minimising the overall cost $C_{\boldsymbol{\alpha}}$.
The vector $\boldsymbol{\alpha}$ expresses the biases in this trade-off game between the cost components. For example, the requirement to speak faster is directly linked to an increase in the value of α_D, which makes shorter versions of a given utterance relatively "cheaper". The cost "saved" this way can be offset against a

proportional increase of one or both of the other two components. The speaker can increase the undershoot (if she's confident that the listener will be able to parse the utterance anyway) or the stiffness (and the force expenditure, for example in a noisy environment). This choice is again reflected in the ratio of the other two weight coefficients α_P and α_H.

The weight coefficients do not prescribe any details of gestural phasing nor, indeed, straightforwardly determine the value of system stiffness; rather, they are high level, intentional parameters of the physically embodied speech production system. It is not our claim that the speakers use a cost components weight distribution (akin to $\boldsymbol{\alpha}$) as parameters of the associated minimisation processes for *online* production of utterances. Rather, we believe that these trade-offs play a vital role during the long-term development of speech as a skilled human activity, and are thus reflected in the phonological laws underlying the speech production (cf. Dispersion-Focalization Theory described in Section 2.1. In other words (paraphrasing Lindblom), speech is adapted to be spoken. During speech acquisition and the accompanying fine tuning of our own production dynamics, we take advantage of these low energy patterns which act as attractors in vastly high dimensional space of all possible productions.

We use a simplified simulated annealing method[2] to identify those model input streams that minimise the overall cost $C_{\boldsymbol{\alpha}}$. The compound function of the pendulum model of the vocal tract and the overall cost function, mapping the

[2]Despite the intended simplicity of our model and cost definition, the objective function $C_{\boldsymbol{\alpha}}$ is still fairly complex with plenty of local minima where simplex (or a hill descent method) tends to "get stuck".

gestural constellations and system stiffness parameters to a single numerical efficiency measure, is used as an objective function of the optimisation problem. For a given model setup, a (non-optimal) initial configuration of input (a collection of gestural activation functions producing the desired utterance and a value of overall stiffness), parameters of the suprasegmental function π and an assignment of cost weight distribution $\boldsymbol{\alpha}$, the optimisation procedure searches the space of possible inputs (adjusting the gestural onset and offset points and the value of overall stiffness) until it finds a gestural constellation/stiffness pair minimising the total cost value $C_{\boldsymbol{\alpha}}$. To guarantee that the optimal input actually produces the given utterance, a very high additional cost is assigned to those inputs which do not produce the required sequence of segments.

Fig. 3 illustrates the optimisation procedure. It shows initial (dashed lines) and optimal (full lines) gestural phasing, computed angular deflections of pendula and resulting prominence functions for the "linguo-palatal" model setup illustrated in Fig. 2A for a gestural sequence ι–$\delta\upsilon$. The initial constellation has been designed so that the subsequent gesture is triggered as soon as the preceding gesture sufficiently approaches its target – reaches a prominence of 0.95 (see the dashed lines on the Activation and Prominence charts). The starting overall stiffness (torsion spring coefficient) of the system was set to $15Nmrad^{-1}$.

These initial input streams are then used as a starting point of the optimisation procedure, along with the model setup parameters and the cost components weight distribution vector $\vec{\alpha} = (1, 60, 100)$. For this particular setup, the (experimentally tested) meaningful values of weight component α_D range

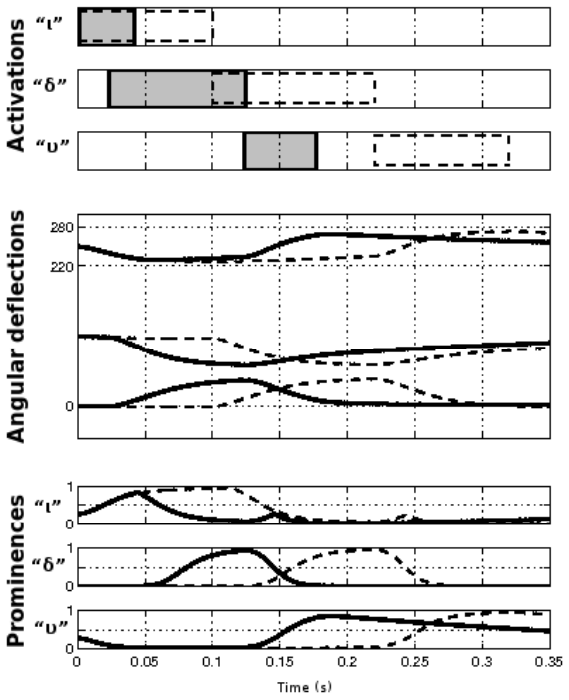

Figure 3: Initial (dashed lines) and optimal (full lines, grey boxes) gestural constellations, and the resulting angular deflections and prominence charts for a sequence ι–$\delta\upsilon$ generated on the model setup illustrated in Fig. 2.

between 20 and 200, the values of α_P range from 10 to 1000.

The full lines show the resulting optimal gestural constellation (grey boxes), the associated angular deflection traces and the prominence charts. The final overall stiffness (not plotted) was calculated as $17.58 Nmrad^{-1}$. Due to the particular details of this model's setup and relatively low durational cost weight, the activation intervals of participating gestures overlap only minimally. The partial overlap for the gestures ι and δ is a consequence of production synergies between these two gestures. The maximal reached prominence of each gesture (vocalic

ones in particular) has decreased for the resulting constellation reflecting the relatively low precision requirements allowing greater undershoot.

4 Simulation Results

Fig. 4 shows a "slice" of the efficient production space for the sequence /ɪ–ðʊ/ produced by the "linguo-palatal" model setup. The top graph shows a sequence of (simplified) optimal gestural constellations (y axis) and an overall trend (interpolation of gestural onsets and offsets) for fixed force expenditure and precision cost weights $\alpha_E = 1$ and $\alpha_P = 100$ and for duration cost weight α_D ranging from 20 to 180 (x axis). The second pane shows the same constellations with their overall duration normalised to see the relative activation durations and overlaps of gestures. (The optimal gestural constellation and the accompanying charts plotted in Fig. 3 show the production details for one instance, that of $\alpha_D = 60$, of Fig. 4.)

Several expected consequences of increasing the weight α_D of the durational cost can be observed in this simulation example: the shortening of the overall duration of the utterance, increase of the overall system's stiffness and increase of target undershoot by participating gestures (not plotted). As mentioned above, this pattern is consistent with human speakers increasing their speaking rate (Ostry et al., 1987; Lindblom, 1963).

More interestingly, the normalised constellation plot captures two less straightforward consequences of speaking rate increase observed by

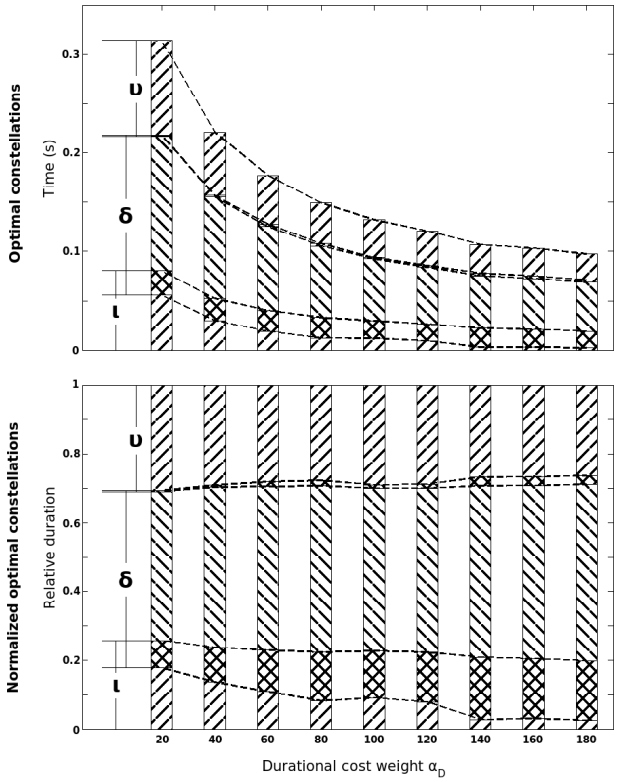

Figure 4: Optimal productions of a sequence ι–$\delta\upsilon$ by the "linguo-palatal" model setup. The top two charts plot a series of the optimal gestural constellations (absolute and normalised gestural activation intervals respectively) for increasing durational cost weight α_D plotted on the x axis. Each bar represents an optimal constellation, its striped boxes correspond to activation intervals of gestures, in a bottom up order, ι, δ and υ, respectively.

phoneticians.

First, it shows that the relative gestural activation durations do not remain constant as the rate increases – the relative duration of the consonantal gesture δ, for example, increases with the rate. This is consistent with Gay's result mentioned in Section 2.1. Moreover, the phasing of gestural onsets and offsets changes in a non-linear fashion: the relative activation overlap of gestures ι and δ gets larger as the overall duration of the utterance shortens. So, our model correctly reproduces the qualitative behaviour of sequencing variations as reported, for example, in Nittrouer et al. (1988) and Nittrouer (1991) and discussed in Section 2.1 of this paper.

Second, the relative phasing of gestures δ and υ which are declared to belong to one suprasegmental unit (syllable) by the durational step function π remains more or less constant with no or little overlap. On the other hand, the activation overlap between gestures ι and δ accross a boundary is more flexible and participates in overall shortening of the utterance with increasing rate – with the cost defined using the function π, to expand the lag between the gestures is "cheaper" across the boundary than it is within a syllable. This is again consistent with observed influence of speaking rate increase on relative gestural timing (see (Cummins, 1999) and our interpretation in Section 2.1).

As mentioned above, when faced with the task of phasing the gestures which impose competing targets on shared articulators, the optimisation procedure generates a gestural constellation with no or small overlaps of the gestures' activation intervals. What happens when subsequent gestures impose less

mutually interacting constraints on the articulators involved in an utterance production; i.e. when the sets of articulators engaged in the subsequent motor actions are comparatively weakly anatomically coupled, as for example in the case of vowels overlaid with bilabial stops? Can we expect an emergence of a strong phasing relationship between similar gestures sharing the common articulators (e.g. vowels) and a separate layer of anatomically quite independent gestures (e.g. bilabial consonants) functionally phased relative to the underlying (vocalic) gestures?

To answer to this question, we used our "linguo-labial" setup described in Section 3 and illustrated in Fig. 2B.

Fig. 5 shows the initial (dashed line) and optimised (full line) constellations for a sequence ι–$\beta\upsilon$ realised in this adapted model setup; the initial overall system's stiffness was set to $15Nmrad^{-1}$, the obtained resulting one was $18.3818Nmrad^{-1}$.

The optimal constellation shows the tight sequential phasing of the vocalic gestures ι–υ (marked by a vertical dotted line and an asterisk) interleaved with the consonantal gesture β. The gesture δ is phased with respect to the vocalic layer according to the *functional* requirement of the correct order of prominence peaks achieved by subsequent gestures – see the prominence panel of Fig. 5. This is an emergent phenomenon: the initial constellation was designed by the same procedure as in the previous case; the gestures were initially phased in a simple sequence[3].

[3]Although Fig. 5 shows only one gestural constellation, the same pattern has emerged for

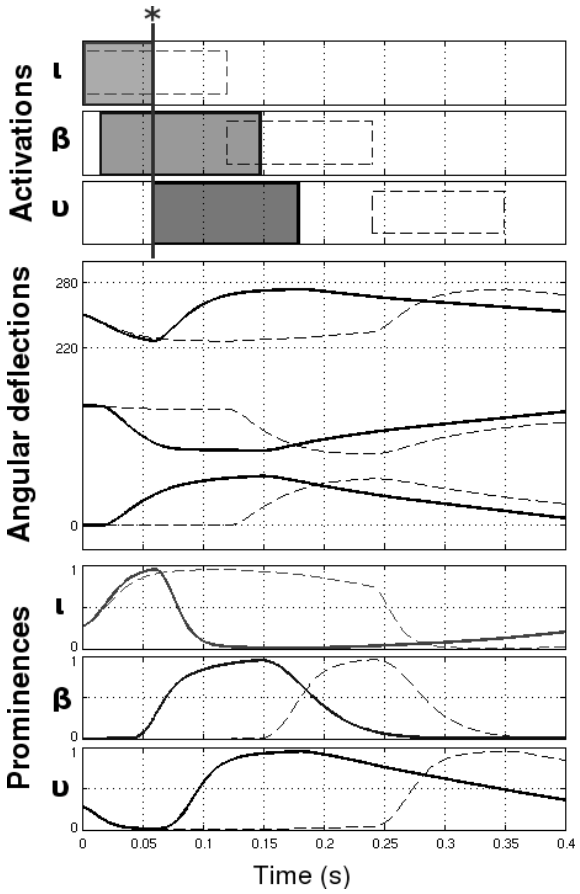

Figure 5: An optimal constellation for a gestural sequence ι–$\beta\upsilon$ by the adapted model setup (see the text) with vocalic attractors relatively weakly anatomically linked with a consonantal (β) attractor. For the description of figure's components, see Fig. 3.

The presence of the two separate layers, one carrying vocalic and the other one consonantal gestures, is analogous to the existence of functional tiers postulated by several phonological theories, mentioned in Section 3; (Fowler, 1983; Browman and Goldstein, 1991), see also (Keating, 1990). Using our modeling approach and, crutially, taking embodiment of the speech production cognitive system seriously, this phenomenon *emerges* as a result of an interaction between sets of constraints that are best captured on two parallel levels of description: the functional constraints formulated on the abstract *task level* and efficiency requirements set down on the embodied *articulatory level*.

5 Discussion

We have argued that some phenomena of speech production and perception that have traditionally been postulated and described in a representational, grammatical fashion by various phonological and phonetic theories (e.g., separation of functional tiers, non-linearity of gestural sequencing with regards to speaking rate, etc.) emerge as consequences of the dynamical properties of a physically instantiated production system, together with the requirement of efficiency in production.

As far as we are aware, our modeling paradigm is the first one to provide a platform for capturing these vital dynamical properties of the speech production cognitive system in a simple and intuitive fashion. It allows us to describe and

an entire series of constellations with varying durational and precision cost weights α_D and α_H similar to one represented in Fig. 4.

explain some of the known production patterns as examples of behaviour of an embodied motor action system, and account for them in the language of intentionally motivated high level parameters linking the system's dynamics, cost functions, and efficiency – without the need of bringing in any additional external phonological postulates. Some of the phenomena, in particular those associated with efficient speech production and sequencing of primitive actions can thus be treated as emergent properties in the sense of Lindblom (Lindblom, 1999).

The intended simplicity of our modeling approach gives an additional support to our claim that it is the character of the task in hand and the *nature* of second order dynamics of an embodied system enforced by the cost efficiency principle (rather then complex details of phonological rules and of vocal tract physiology) that on their own bring about important phenomena accompanying speech production.

We do not claim that an account of an embodied articulatory system is the only "correct" level of description on which it is possible to talk meaningfully about phonological and phonetic phenomena. We agree, for example, with the school of AP that there are at least three such informative levels, each one designed to best describe constraints imposed on speech production by its linguistic, functional and physiological aspects, respectively. Identifying any of these constraints and, as seen in the second example in the previous section, the cross-level interdependencies between, them helps us to better understand the cognitive processes underlying the production and perception of speech.

The model setups presented in this paper reflect only very high level organisational characteristics of the end effectors of the human vocal tract. In

follow up work, we have studied models that more closely capture the anatomical and functional properties of the speech production system. In these models, the task-oriented and embodied aspects of model articulators are defined hierarchically at interconnected levels of description closely related to the architecture used in the task dynamical implementation of AP (Saltzman and Munhall, 1989). In order to be able to use these models of the vocal tract in conjunction with the optimality paradigm presented here, we adapted the task dynamical definition of their behaviour so that it takes into account the embodied character of speech articulation.

The results of simulations performed on these instances of the general abstract modeling paradigm presented here provide further insights into emergent patterns of coarticulation and functional tier separation. Moreover, as reported elsewhere (Simko and Cummins, 2009), the details of the optimal articulatory behaviour generated by these models are in a considerable agreement with the gestural sequencing patterns manifested by human speakers (Browman and Goldstein, 1988; Löfqvist and Gracco, 1997; Löfqvist and Gracco, 1999; Löfqvist and Gracco, 2002).

This article has presented a novel mathematical model in some detail. It is, perhaps, worthwhile to summarise the overall account provided herein. We begin with many of the same assumptions as underwrite the Articulatory Phonological framework and its task dynamic implementation: we assume that gestures are primitives of linguistic organisation, that they are crucially sequenced in time, and that their evolution is constrained in lawful fashion by a task-specific

dynamic regime. In contrast to previous approaches, we choose to define tasks in the space of physically instantiated articulators, and we make use of the inertial properties of these articulators to shed light on the sequencing of gestures in real time, and in dependence on such high-level speech properties as speaking rate. We find that it is possible to operationalise the conflicting constraints of articulatory ease and perceptual clarity within a single additive cost function. We then demonstrate, through simulation, that this cost function, and the overarching concept of efficiency, can indeed suitably constrain the organisation of gestures in time.

References

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

Barry, W. J. (1998). Time as a factor in the acoustic variation of schwa. In *ICSLP-1998*, Sydney, Australia.

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Ph.D. dissertation, University of Amsterdam.

Browman, C. P. and Goldstein, L. (1991). Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J. and Beckman, M. E., editors, *Papers in Laboratory Phonology I: Between the Grammar and the*

Physics of Speech, pages 341–376. Cambridge, U. K.: Cambridge University Press.

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

Browman, C. P. and Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlee*, 5:25–34.

Browman, C. P. and Goldstein, L. M. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, 45:140–155.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, 24:209–244.

Byrd, D., Kaun, A., Narayanan, S., and Saltzman, E. (2000). Phrasal signatures in articulation. In Broe, M. and Pierrehumbert, J., editors, *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, pages 0–87. Cambridge University Press, London.

Byrd, D. and Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2):149–180.

Cooke, J. (1979). The organization of simple, skilled movements. In *Motor Learning and Control*, pages 199–212. NATO Advanced Study Institute, Senan-que, France.

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

Feldman, A. and Latash, M. (2005). Testing hypotheses and the advancement of science: recent attempts to falsify the equilibrium point hypothesis. *Experimental Brain Research*, 161(1):91–103.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in sequences of monosyllabic stress feet. *Journal of Experimental Psychology: General*, 112:386–412.

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

Gay, T., Ushijima, T., Hirose, H., and Cooper, F. S. (1974). Effects of speaking rate on labial consonant-vowel articulation. *Journal of Phonetics*, 2:47–63.

Guenther, F. H. (1995). Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production. *Psychological review*, 102(3):594–621.

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

Howard, I. S. and Huckvale, M. A. (2005). Training a vocal tract synthesizer to

imitate speech using distal supervised learning. In *Proc. SPECOM 2005*, pages 159–162, Patras, Greece.

Iskarous, K., Goldstein, L. M., Whalen, D., Tiede, M., and Rubin, P. (2003). Casy: the haskins configurable artiticulatory synthesizer. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain. Universitat Autònoma de Barcelona.

Keating, P. A. (1990). The window model of coarticulation: articulatory evidence. In Kingston, J. and Beckman, M., editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 451–470. Cambridge University Press.

Kelso, J. A. S. (1995). *Dynamic patterns: The Self-Organization of Brain and Behavior*. MIT Press, Cambridge, Massachusetts, London, England.

Latash, M. (2008). *Synergy*. Oxford University Press.

Liljencrants, J. and Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862.

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *Journal of the Acoustical Society of America*, 35(11):1773–1781.

Lindblom, B. (1983). Economy of Speech Gestures. In MacNeilage, P. F., editor, *The Production of Speech*. Springer-Verlag, New York.

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

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

Lindblom, B. (2000). Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica*, 57(2-4):297–314.

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

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

Löfqvist, A. and Gracco, V. L. (2002). Control of oral closure in lingual stop consonant production. *Journal of the Acoustical Society of America*, 111(6):2811–2827.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication*, 1:199–229.

Nittrouer, S. (1991). Phase relations of jaw and tongue tip movements in the

production of VCV utterances. *Journal of the Acoustical Society of America*, 90(4):1806–15.

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

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

Ostry, D. J. (1986). On viewing motor behavior as a physical system. *Journal of Phonetics*, 14:145–147.

Ostry, D. J., Cooke, J. D., and Munhall, K. G. (1987). Velocity curves of human arm and speech movements. *Experimental Brain Research*, 68:37–46.

Ostry, D. J. and Feldman, A. (2003). A critical evaluation of the force control hypothesis in motor control. *Experimental Brain Research*, 221:275–288.

Perrier, P., Perkell, J., Payan, Y., Zandipour, M., Guenther, F., and Khalighi, A. (2000). Degrees of freedom of tongue movements in speech may be constrained by biomechanics. In *Proceedings of the 6th International Conference on Spoken Language Processing, ICSLP 2000*, Beijing, China.

Saltzman, E. L. (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. L. and Kelso, J. A. S. (1987). Skilled Actions: A Task-Dynamic Approach. *Psychological Review*, 94(4):84–106.

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

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

Simko, J. and Cummins, F. (2009). Sequencing of articulatory gestures using cost optimization. In *Proceedings of the Interspeech Conference*, Brighton, UK.

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

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