As time goes by:
A critical appraisal of space and time in Articulatory Phonology in the 21st century

Khalil Iskarous
Department of Linguistics University of California at Los Angeles, USA; kiskarou@usc.edu

Marianne Pouplier[1]
Institut für Phonetik und Sprachverarbeitung, LMU Munich, Germany;
pouplier@phonetik.uni-muenchen.de

Running head: A critical appraisal of space and time in Articulatory Phonology

Corresponding Author
Marianne Pouplier
Institute of Phonetics and Speech Processing
LMU
Schellingstr. 3
80799 München, Germany
pouplier@phonetik.uni-muenchen.de
+49-89-2180-2811

Keywords: Articulatory Phonology, Task Dynamics, dynamical systems, π-gesture, syllable, prosody, planning

[1] The authors are in alphabetical order.

Abstract

Articulatory Phonology and Task Dynamics have modelled spoken language mathematically based on dynamical systems, expressing the view that speaking is contiguous in nature with many other biological phenomena that can be described in this way. In this paper, we present a critical appraisal of the developments in Articulatory Phonology and Task Dynamics in the 21st century. Our paper identifies some of the key areas in which progress has been made, and others in which more progress is warranted. We thereby touch on recent work contributing to the empirical underpinning of some assumptions of the Task Dynamic model. We further consider recent proposals of how Articulatory Phonology can deal with linguistically structured macro- and microscopic variation in constriction gestures induced by syllabic and phrasal prosodic structure. Part and parcel of these developments is the integration of the dynamical expression of phonological contrast into a model of utterance planning. Central to our exposition is the integration of accounts for contrastive spatial goals with accounts for the prosodic structuring of time. We finish our overview with a discussion on how a stronger link between articulation and acoustics could further enhance the dynamical approach to spoken language.

Highlights

- Recent developments in Task Dynamics and Articulatory Phonology are critically exposed, including
- Proposals for time-varying versus nonlinear stiffness
- Prosodic modulation gestures
- Models of planning dynamics

1. Introduction

Articulatory Phonology is a framework that attempts to answer two central questions at the basis of Linguistic Phonetics: For one, how do the planning and performance of language integrate into one skill of linguistic speech production? Secondly, how do the empirical generalizations regrading complex spatiotemporal phonetic signals follow from the linguistic structuring of speech? Articulatory Phonology attempts to answer both questions for the study of spoken language[2] through the idea that language provides the speech motor control system with a set of spatial or temporal *goals*, which could be defined at various levels: a constriction gesture, syllable structure, or prosodic boundaries. The many intricate details of phonetics then drop out of how these goals are achieved in a variable, yet organized, fashion. What sets Articulatory Phonology apart from other approaches to spoken language is the aim to integrate the cognitive and physical aspects of speech into one unified theory by defining these goals based on the mathematics of dynamical systems (among others, Fowler et al., 1980; Browman & Goldstein, 1986, 1989; Saltzman & Munhall, 1989; Byrd & Saltzman, 2003; Tilsen 2019; Sorenson & Gafos, 2016). The reason that dynamical systems, invented to explain the movement of celestial objects and developed further to describe many aspects of the natural world, are useful for describing speech is that the cognitive concept of a discrete linguistic goal can be mapped to the dynamical concept of a *stable equilibrium* or *attractor*. Many systems in nature have an equilibrium position, to which they return, or are attracted to, if perturbed. As an example, consider the viscoelastic tissue on your forearm; if you pull, it goes back, and if you push it, it goes back to its static attractor state. Heart pumping at a particular rate is a dynamic attractor goal that usually resists all kinds of perturbations, and the mathematics of dynamical systems is used to formalize these behaviors. Inspired by dynamical theoretical accounts of many complex phenomena in nature, the proponents of Articulatory Phonology have sought to apply dynamical systems theory to model spoken linguistic communication in a non-dualistic fashion, seeking to overcome the traditional split between the disciplines of phonology and phonetics.

Articulatory Phonology has evolved to have three major interconnected dynamical components which model speech as goal-seeking behavior at various levels: 1) contrast dynamics, focusing on how linguistic contrasts within a word are identified with the *spatial* goals, predicting in which way the actual production of words can be variable within and across languages; 2) prosodic dynamics, focusing on how a language's syllabic, prominence, and grouping systems structure the *timing* of contrastive goals in word and phrase production; 3) planning dynamics, focusing on the temporal unfolding of utterance planning and feedback.

The aim of this current paper, in line with the theme of the special issue, is to render a critical appraisal of the developments this particular theoretical approach has taken since the turn of the millennium. To that effect, we will critically expose the contrast, prosody, and planning components of the theory, with an emphasis on the centrality of time to understanding linguistic phonetics, and we will point out places where additional theory needs to extend and connect existing theories within the framework. For readers not thoroughly familiar with Articulatory

[2] Even though most work in the framework has focused on spoken language, some work has started to look at signed language (Tyrone et al. 2010).

Phonology, we provide brief conceptual overviews of central aspects to allow for full appreciation of our review of recent themes and issues, some of it in Appendices. Detailed introductory overviews can be found, for instance, in Gafos & Goldstein (2012), Pouplier (2011; 2020), Goldstein et al. (2006), Fowler & Iskarous (2012), and Iskarous (2017).

There have been many insightful studies conducted within the Articulatory Phonology framework since the dawn of the 21st century, and we do not have the space to discuss them all in detail. The main focus here will be on recent contributions to what we believe to be one of the largest novel conceptual proposals since the original development of Task Dynamics, and this concerns the question of how language structures temporal macro- and microscopic variation in constriction gestures at the levels of syllabic and phrasal prosodic structure. Part and parcel of this is the integration of the contrast dynamics part, which was at the heart of the first exposition of the theory, into a model of utterance planning. Prior to this, we will review recent developments in the contrast dynamics component which has raised new perspectives on how gestures should be defined mathematically and by which principles overlapping gestures interact. Highly interesting other developments, such as the incorporation of tonal systems and pitch gestures (Gao, 2009; Karlin, 2018), the split-gesture dynamics proposed by Nam (2007), or the FACTS model of speech motor control (Parrell et al, 2019) although by no means of lesser importance, will for space reasons not be covered in detail here.

2. Contrast and Spatial Goals

We will start with new extensions to Task Dynamics, the 1989 theory of how contrasts are encoded dynamically (Saltzman & Munhall, 1989). Since the extensions necessitate some understanding of how the original theory works, we start with some general background to Task Dynamics and we additionally have provided Appendix A, which gives a conceptual understanding of what a differential equation or dynamical system is, and what the central equation of Task Dynamics means.

2.1. Goals in Task Dynamics

Three developments in the scientific understanding of cognition paved the way for Task Dynamics and the particular way in which it seeks to integrate phonology and phonetics. The first development is the hypothesis that brain systems, and cognition itself, are continuous with the rest of nature, and therefore can use the same type of mathematical analysis as had been used before, differential equations. This is the view of Grossberg (1968, 1972, 1978), Wilson & Cowan (1973), and Amari (1977). And it is this view that made it felicitous to take the tools that describe celestial orbits, electromagnetic waves, and tissue morphogenesis and apply them to the cognitive process of phonological contrast. The second development, which occurred in the 1970's, is the philosophically and psychologically surprising finding that the human brain's inner structure and circuits are highly reflective of the structure of the world. For instance, Goldman-Rakic (1987) found that working-memory prefrontal circuits can maintain events in their temporal order and spatially in a way that reflected their layout in space, O'Keefe & Dostrovsky (1971), in Nobel-prize winning work, found place cells in the hippocampus that become active when an animal is in a particular location (O'Keefe & Nadel, 1978), and

Georgopoulos et al. (1986) found motor cortex cell populations that fire depending of the direction of reaching an animal performs. These results pointed to the degree that the structure of brain circuits that serve cognitive and movement tasks actually reflect the structure of the outside world. This then newly discovered parity of brain and world made it at least possible that something as cognitive as phonology could be related to the physiological system in the world that actually realizes it, articulation, making it possible to think of an *articulatory phonology* at all. The third development is the realization by Bernstein (1947), and later by Greene et al. (1972), as well as Turvey (1977) that animal movement is a hierarchical process where a higher-order goal is accomplished through the synergy of effectors, and that this can be studied using a dynamical systems approach in which the goals are equilibria that the system converges to, independently of the initial conditions or perturbations to its trajectory. (Saltzman, 1979; Fowler et al., 1980; Ostry & Munhall, 1985). And it is this hierarchical dynamical approach to movement that allows the integration of an abstract movement goal (for example, lip closure) and the synergy of effectors implementing the movement towards the goal (for example, upper and lower lip, jaw) into a single description. There are two main parts to Task Dynamics, which will be discussed in turn, each with its extensions since 1989. The first part is the dynamical behavior of tasks, and the second is how task changes organize articulator changes.

As discussed in detail in Appendix A, Task Dynamics proposes that a word consists of several *task variables*, called *gestures*, most describing a Constriction Location or Constriction Degree goal for a particular vocal organ (e.g., tongue tip constriction location/degree or glottal aperture), with each gesture being described mathematically by a critically damped second-order dynamical linear system. The contrastive units of phonology are therefore defined with reference to spatio-temporal parameters. The constants of these dynamical systems are for one, the *target*, which specifies the constriction location or degree of a task/linguistic goal, and secondly, the *stiffness* which specifies how fast the equilibrium value/goal is achieved. It is in these constant values that speech sounds contrast, as both parameter values are set by the phonological system of the language. For instance, target constants for Lip Aperture (LA) for a labial stop and a labial fricative will be different, and stiffness values for consonants and vowels will be different, as the goals for the former are achieved faster.[3] The solution to the differential equation for the task variable is then the phonetic task change, for instance the change in LA as a stop is approached in /ap/. The central equation of Task Dynamics, for an LA gesture as an example, can be written as:

$$(1) \qquad LA_{tt} + 2LA_t + k\,(LA - LA_0) = 0,$$

where LA_{tt} is the acceleration, LA_t is the velocity, LA the current value of lip aperture, and LA_0 the goal (equilibrium position). The coefficient of velocity is 2, as that insures critical damping, as explained in Appendix A.

This equation relates the acceleration, velocity, and current value of LA with the constants *k* and LA_0 for the particular vowel or constant and in this way integrates phonological constants and time-varying phonetic change of tasks into one description: The constants of the equation

[3] Prosodic variation has sometimes also been linked to the stiffness parameter (see Section 4).

are invariant while the given gesture is active, while they simultaneously encode how the gesture evolves in space and time. The timing and duration of these gestures was set by hand initially (but see later discussion of the syllable model, as well as the "Serial Dynamics" section of Saltzman & Munhall, 1989).

Since equation (1) applies at every point in time, meaning that whatever the acceleration, velocity, and value of LA are, their values as combined by k and LA_0 must sum to 0, the predictions of this mathematical statement are rigorously empirically testable. Indeed, quite early in the development of the theory, Perrier et al. (1988) noted that the critically damped system's peak velocity is predicted to be achieved much earlier than it does in real movement. Figure 1a shows the time change in the LA task variable (top) and its velocity (bottom) in changing from /a/ to /p/ in a natural production (obtained from the Euclidean distance of the upper and lower lip sensors in articulography data), while Figure 1b shows the 2nd order simulation of that motion. The early achievement of the velocity maximum was therefore seen as a failure of the predictions of the critically-damped second-order linear dynamical system model for tasks.

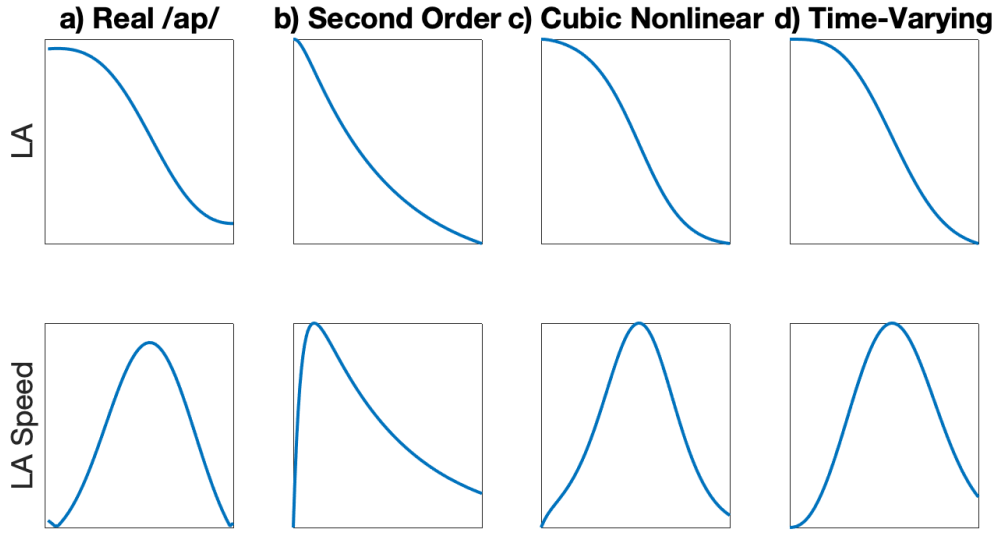

Figure 1. Comparison of lip aperture (top) and velocity magnitude (bottom) of natural data (a), with simulations of a linear second order system (b), a system with time-varying stiffness (c), and nonlinear stiffness (d). The x-axis depicts time in all cases.

There have been two approaches to modifying the dynamical model to delay the time at which the magnitude of velocity (speed) peaks, one of them older and one of them fairly recent. In the first, Perrier et al. (1988) proposed a solution which was further developed by Kröger et al. (1995) and Byrd & Saltzman (1998) where the stiffness k is no longer a constant, but varies in time. The consequence is a gradual waxing and waning of gestures at their on- and offset, instead of gestures being turned on and off as step-functions. This means that a gesture's activation strength gradually rises and falls at its edges. Introducing time-varying stiffness leads to a situation of so-called *non-autonomy* in dynamical systems theory. Non-autonomy refers to the system's behavior being no longer exclusively due to the dependence of the velocity on the

position and acceleration of the system, and constants, but also on some externally imposed time-variation. In the non-autonomous account, *k* starts small, then increases, then decreases again. If *k* starts out very small, there will hardly be any change, then if *k* rises, there will be a sharp change, and if *k* decreases again, then we will be a slowing of change. This is exactly what happens in the natural data of /ap/ in Figure 1a: slow decrease, sharp decrease, then slow decrease. Therefore, the concept of stiffness increasing and decreasing over time was introduced. The specific model of Byrd & Saltzman (1998) predicts the change in Figure 1d which is highly reflective of the real change in Figure 1a.

The second approach, proposed recently by Sorenson & Gafos (2016) is, effectively, to make *k* depend not on time but on the task variable state itself and its distance to the target: if there is a large or small difference between the task state and the target, *k* will be small, and there is an intermediate difference, *k* will be large. Again, this achieves the desired result as seen in the simulation in Figure 1c. This is mathematically achieved by adding a term to the differential equation that is proportional to the cube of the task variable. The details can be found in Sorenson & Gafos (2016). This nonlinear approach is autonomous, since no outside factor determines the evolution of *k*. The latter solely depends on the state itself, making the system self-dependent. There has as of yet not been much of an empirical reason yet to decide between the two versions of the differential equation that yield a delayed velocity peak. But as Sorenson & Gafos (2016) argue, it is always preferrable to assume an autonomous dynamical system, since no recourse to some external time-varying force is necessary. When we discuss the dynamic modelling of prosody in terms of the π-gesture in Section 4, it will turn out that the time-varying nature of the gesture's activation is an important aspect of the model, with profound implications for the decrease in overlap of contrastive gestures observed at prosodic boundaries. Extending Sorensen & Gafos' (2016) proposal for contrastive gestures to the prosodic level has yet to be accomplished.

2.2. Synergy in Task Dynamics

Saltzman & Munhall (1989) also proposed a mathematical relationship between each task variable velocity as a function of time and each articulator as a function of time. This relationship specifies how certain articulators synergize their efforts to achieve a task variable change. These synergies have also been called *coordinative structures* (Turvey, 1990). For instance, here is the relationship for lip aperture:

(2) $LA_t = w^{LA\_UL}\, UL_t + w^{LA\_LL}\, LL_t + w^{LA\_JW}\, JW_t$.

which means that the velocity of change of the task variable LA is linearly related to the velocities of the Upper Lip (UL), Lower Lip (LL), and Jaw (JW), with given linear coefficients called weights *w*. These latter coefficients specify which articulators contribute the most to a particular tract variable 's achievement at any given instance. These coefficients incorporate the important concept of *motor equivalence* into the model. For instance, if the lower lip is observed to move the most for changing LA, the upper lip a bit less, with the jaw moving least, the weight for JW will be highest, the UL will be intermediate, and LL will have the smallest *w*. The weight coefficient may change flexibly depending on context. The total set of such relations mapping articulator velocities onto task variable velocities is called the *Jacobian*.

Basic techniques form Calculus (chain rule) can then be used to the generation of differential equations for the velocities of articulators from the differential equations for the task variables. These equations can be solved to yield predictions about how the articulators will actually move, making the theory, again, thoroughly testable. These solutions will depend on both the phonologically-determined k and x_0 for each task as well as the weights w, specifying the synergy. We would like to note here that it is in these equations for LL_t, UL_t, JW_t, etc., and their solutions that the unification of phonology and phonetics is achieved the most: phonological constants determine the phonetic movements of the movements of the articulators, but these phonological constants have reality only in how they actually structure the phonetic movements through the dynamical system. Therefore this is a model where phonetics and phonology define each other.

The weights were initially set by hand so that when the articulator differential equations are solved, the predicted articulator trajectories and simulated acoustics were qualitatively as natural as possible. Two strands of recent work have demonstrated how synergies can be inferred numerically from empirical data. For one, Iskarous et al. (2013) used the concept of *mutual information* (MI) from Information Theory to calculate the influence of different articulators on different task variables directly from data. X-ray microbeam data of speakers producing CV syllables were used to determine for each consonant, the degree to which its achievement necessitates each articulator. This was done by fixing the consonant, and varying the vowel, and measuring for each articulator, how well one can predict the articulator's position in the middle of the consonant from the articulator position in the middle of the vowel. This predictability was quantified through MI. If a consonant, e.g., /p/, has an LA task it was found that the tongue back position in the middle of the consonant was highly predictable from the position of the tongue back for the vowel, with MI quite high. But, the jaw and lower lip positions during the consonant have low MI with respect to the vowel, as the LA task for /p/ recruits these articulators, making them only limitedly available during /p/ to anticipate the articulation of the following vowel. Iskarous et al. (2013) therefore suggested that the weighting coefficients can be estimated from the MI values, precisely because these contain information about which articulators contribute the most for each task. Chen et al. (2015) extended the MI framework to fixing the consonant and varying the vowel to learn about the contributions of articulators to mostly vocalic tasks, and Abakarova et al. (2018) used the MI idea to study the development of coarticulation from children to adults, offering hope that this framework could lead to a deeper understanding of how synergies are learnt.

Secondly, Lammert et al. (2012) used artificial neural networks and locally-weighted regression to estimate the entire Jacobian from solutions of the differential equations discussed earlier, and then applied them to real time MRI data, showing that it is possible to estimate synergies directly from data. Using factor analysis, Sorenson et al. (2019) went further in estimating synergies from real time MRI data, leading to greater understanding, for instance, in the degree to which the jaw helps the accomplishment of different tasks. This study has also probed the degree to which synergies are speaker-independent, but there are still no clear results on this matter.

One important issue on which there has not been much progress is how to model articulators and how to complement Task Dynamics with an actual biomechanical model.[4] So far, the computationally implemented Task Dynamic model (TaDA) has used the Configurable Articulatory Synthesizer Model (Nam et al. 2004), which models articulators as mostly rigid geometric curves. This model has been quite useful due to its simplicity, but we know that most of the speech articulators, tongue, lips, and velum are extremely flexible muscular hydrostats. Therefore, if predictions from the Task Dynamic model are to be compared with articulatory data, especially from ultrasound and MRI, which capture a lot of this flexibility, there an actual biomechanical model of articulator motion is required. Such a model is available in the Artisynth muscle model of speech articulators (Lloyd et al., 2012, see also Buchaillard et al., 2006), but we are not aware of attempts to implement Task Dynamics using this flexible model of speech muscle deformation.

2.3. Gestural Overlap

Browman & Goldstein (1989) proposed that linguistic contrasts (tasks) can occur at the same time, and therefore overlap. That contrasts are not purely sequential, but possibly parallel, had already been proposed within the framework of Autosegmental Phonology (Goldsmith, 1990; Clements, 1976). Articulatory Phonology brought time and space into phonology in a much deeper way, though, as it considered many local phonological alternations to be due to the sliding of gestures past one another, resulting in gestural blending and gestural hiding (Browman & Goldstein, 1995). One of the earliest arguments for continuous time in phonology, not just in the executed phonetic signals, is what Steriade (1989) called Dorsey 's Law, a phenomenon that occurs in many languages where a vowel is epenthesized, and has the same quality as a contiguous vowel in the word. Steriade argued that the temporally short consonantal gesture simply slides along a temporally longer vowel gesture, resulting in portions of the vowel gesture showing up on both sides of the sliding consonantal gesture. This is evidence that temporal fluidity is in the phonology itself, not just in an articulatory execution. This type of explanation would not be available in a theory with a *phonology-extrinsic timing*, as that proposed recently by Turk and Shattuck-Hufnagel (2021), since it is the manipulation

[4] Šimko & Cummins (2010) suggested an "embodied Task Dynamics" that incorporates effector-specific cost functions for articulatory and perceptual effort at the task level, yet their model does not incorporate actual articulator biomechanics.

of spatial goals in time *within the phonology* that is at the heart of the explanation.[5] This fluidity of temporal and spatial goals was argued to be at the basis of many synchronic and diachronic alternations in phonology (Browman & Goldstein, 1989, 1991). There has for instance been considerable attention on how assimilation at word boundaries may be accounted for based on the temporal sliding of gestures (e.g. Pouplier et al., 2011; Pouplier & Hoole, 2016; Son et al., 2007).

Saltzman & Munhall (1989) proposed that when two gestures overlap, they have joint control of the vocal tract. In case they impose conflicting demands on the same articulators, their influence is blended. They proposed additional constants (weights for the blending parameters) in the model that determine which of the two gestures exerts greater control of the vocal tract. Iskarous et al. (2012) proposed that the blending principles in the case of overlap should be assumed to be language-specific which could account for differences in coarticulation between consonant and vowels in Navajo, Spanish, and English. They argued for treating the task variables constriction location and degree independently of each other in terms of blending as a way to account for language-specific vowel-induced consonantal variation. For English, velars undergo a pronounced change in place of articulation with vowel context, but constriction degree is resistant to coarticulation, since velar closure is invariably achieved. In Navajo, the velar fricative /x/ shows extreme variation in both constriction location and degree as a function of the following vowel. Iskarous and colleagues use the computational task dynamic synthesizer to show that a blending of the constriction location to equal parts between vowel and consonant gives rise to the English velar fronting effect. For constriction degree, an asymmetric blending with a complete dominance of the constriction degree parameter of the consonant is assumed – causing closure to be invariably achieved independently of vowel context. In Navajo, both constriction location and degree are averaged between consonant and vowel in the case of conflict, resulting in a blended place of articulation *as well as* a blended aperture value. Spanish uses yet a third pattern in which constriction degree, but not location is blended, leading to spirantization. This is an example of one of the few studies that have sought to apply the concept of spatial overlap built into Task Dynamics to account for cross-linguistic variation.

---

[5] In their argument against the spatiotemporal phonology of Articulatory Phonology/Task Dynamics, Turk & Shattuck-Hufnagel (2020) have favored an autonomous symbolic phonology, which we understand to mean a module abiding by cognitive-linguistic laws quite separate from the spatiotemporal concerns of phonetic and motor-sensory modules. However, the incorporation of deep phonetic spatiotemporal notions into phonological representation is not unique to Articulatory Phonology. For instance, Chomsky & Halle (1968) in Chapter 9 of SPE argued that abstract phonology is too powerful if not phonetically constrained in that it predicts many phenomena that could never exist in the world's languages, and proposed a markedness theory that refers to the structure of the phonetic systems. Similarly, the authors of Optimality Theory have argued for the phonetic basis of one of the two types of phonological constraints: "If Phonology is the computational link between the lexicon and phonetic form, then Markedness is the advocate of the phonetic interface, Faithfulness the agent of the lexical interface" (Prince & Smolensky, 2003, p.2). Therefore, the assumption of a phonetics-free symbolic phonology is a difficult one, given the history of phonology itself.

3. Timing and the Syllable: Prosodic Planning Dynamics I

The original proposal of Articulatory Phonology was focused almost exclusively on the substance of contrastive phonological representations, but soon after, attention shifted to higher linguistic structure, focusing on how the basic contrastive representations – often metaphorically termed the *atoms* of speech production – can in principle be grouped into a larger *molecular* unit, the syllable. Browman & Goldstein (1995) contended that syllabic structure emerges from particular timing, or coupling, relations between gestures. It was empirically observed that the relative timing of prevocalic and postovalic consonants relative to the vowel is quite different, and it was concluded that it is this difference that gives rise to what is traditionally considered syllabic structure (Browman & Goldstein, 1988). Browman & Goldstein (2000) expanded this into an actual dynamical model of the syllable, sometimes also referred to as *coupled oscillator model of syllable structure*. This was also the first major model of planning dynamics, an area we will cover in detail in Section 5.

Modelling syllable structure in this view requires a dynamical system in which precise timing and delay of one event with respect to another are at the center of the description. The main such time-keeping system in mathematical biology which was also adopted in Articulatory Phonology is the *limit cycle*. Each gesture is assigned its own such limit cycle. We have devoted Appendix B to provide physical intuition on what a limit cycle oscillator is, how it works, and how limit cycles couple to each other. We would like to emphasize here that even though mass-spring systems are often used to describe how a limit cycle works, the limit cycle is not used as a source of spatial *movement* in Articulatory Phonology's account of syllable structure, but as a time-keeping and time-delay mechanism. In this aspect of the theory, Articulatory Phonology actually separates time and space, a theoretical feature that also Turk & Shattuck-Hufnagel (2020) advocate and provide evidence for, but that we will later criticize.

Since each gesture has an oscillator, grouping effects can be modelled via coupling of oscillators. The basic proposal therefore is that atomic gestures cohere via local, pairwise coupling relations into larger molecular structures such as segments[6] or syllable onsets. This means that the syllable is viewed as being the expression of particular levels of cohesiveness between pairs of gestures as coupled oscillators. Phase relationships specify the specific nature of gestural cohesion in terms of relative timing. By hypothesis, pairs of gestures are coupled locally to one another either in-phase (0°, synchronously) or anti-phase (180° out of phase). Hereby Articulatory Phonology has relied heavily on the fact that in-phase synchronization is in the general movement sciences as much as in the science of synchronization known to have a special status as a natural primitive – 0° synchronization is maximally stable to perturbation and does not have to be learnt (Turvey, 1990). Browman & Goldstein thus proposed that CV

[6] Note that originally the segment was not recognized as an independent unit by Browman & Goldstein (1986) – segments, they proposed, are constellations of multiple gestures, no different from e.g., CV. In that sense, every consonant or vowel requiring more than one gesture is a cluster or complex molecule. This was under some debate when Articulatory Phonology was first proposed (Saltzman & Munhall, 1989; Byrd, 1996; Byrd et al., 2009), but has received little attention since (but see Fowler & Goldstein, 2003). The 'grain size' of the atomic units has, however, been revised in Nam's (2007) split-gesture dynamics, in which constriction formation and release gestures for consonants (not vowels) are modelled as independent gestures which can enter distinct coupling relations.

sequences, known to be a linguistic primitive, are such a universally abundant structure in spoken language because consonant and vowel gestures are coordinated 0° phase. This assumption has been bolstered empirically by reports that movement initiation of consonants and vowels in #CV is within less than 50ms of each other (Löfqvist & Gracco, 1999; Gubian et al., 2019; Liu et al., 2022).

3.1. The coupled oscillator model of syllable structure and the c-center effect

In the first expositions of Articulatory Phonology, gestural scores, which specify the gestural composition of an utterance and the relative timing of the individual gestures, were simply given. Browman & Goldstein (2000) extended their basic proposal of how gestures cohere into larger units into an actual dynamical model of the syllable, also allowing for the possibility of complex syllable margins. This was achieved by allowing for incompatible, competing phase relationships between pairs of gestures. This proposal essentially introduced for the first time a notion of planning within the gestural model: The concept of competing phase relationships presupposes computational planning time which corresponds to the settling time of the competing oscillators (more on this below in Section 5): Lexically specified, conflicting phase relationships will result in the computation of a weighted output, just as is the case in the Task Dynamic model when gestures impose competing spatial demands on the same articulator. Competing phase relationships are hypothesized to be a unique property of prevocalic consonants, i.e., of syllable onsets. This has become known as the *c-center effect*.

The c-center is a diagnostic tool on how to apply timing measurements to kinematic data in order to uncover underlying competing phase relationships and hence syllabic structure. By hypothesis, all consonants of an onset are coupled in-phase to the vowel, but anti-phase to each other. In the case of multiple onset consonants, this leads to incompatible phase relationships that cannot all be satisfied. The result is observable at the kinematic level as the c-center. This is illustrated schematically in Figure 2.

For instance, in Figure 2 (right), it can be seen that the gesture for the consonant adjacent to the vowel shifts relative to the anchor (marked by a solid line) as a function of increasing onset complexity, whereas the temporal midpoint of the onset as a whole (the c-center) is unchanged in its timing to the anchor, as predicted by the model laid out in Browman & Goldstein (2000). At the level of the surface kinematics, this leads in an C1C2V structure to increasing C2V overlap compared to singleton C2V. This stability is result of the higher number of pairwise, local consonant-to-vowel coupling relations in onsets compared to codas. In coda there is by hypothesis only sequential coupling: V-C1-C2, and hence V-C1 timing is unaffected by increasing coda complexity.

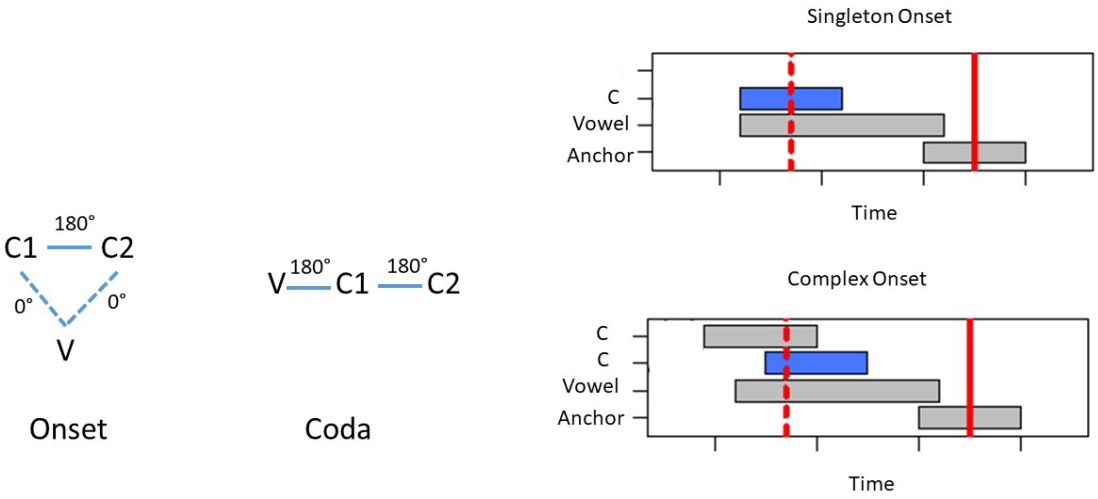

Figure 2. Left: Coupling graphs illustrating the coupled oscillator model of syllable structure. Right: Schematic illustration of the the c-center effect. Observe how the consonant immediately adjacent to the vowel increases its overlap with the vowel as onset complexity increases from C to CC. The dashed line indicates the temporal midpoint of the onset corresponding to the c-center. The solid line is the anchor relative to which timing is evaluated.

The coupled oscillator model of syllable structure and the idea that syllabic structure is expressed articulatorily in differential timing relations has received a considerable amount of empirical testing across a range of languages (among others, Gafos et al., 2013; Shaw et al., 2009; Hermes et al., 2013; Pouplier & Beňuš, 2011; Pastätter & Pouplier, 2017; Marin & Pouplier, 2012; Sotiropoulou et al., 2020; Brunner et al., 2014). A common heuristic for identifying the hypothesized underlying coordination relationship has been to investigate the relative timing of various articulatory landmarks, such as the temporal midpoint of the onset (whether simplex or complex) or of the vowel-adjacent consonant change relative to a fixed anchorpoint later in the word (Figure 2). Also the relative stability (in terms of standard deviation or coefficient of correlation) of these temporal landmarks across conditions has been used to index the underlying coupling structure.

Overall this research has been able to confirm that differences in relative timing between consonants and between consonants and the vowel express syllable structure, yet the empirical results have also uncovered a rather complex picture. For one, onsets and codas exhibit different overlap patterns between successive consonants, as already discussed in Byrd (1996; see also Marin & Pouplier, 2014; Shaw & Gafos, 2015): onset consonants are less overlapped with each other compared to codas, yet this does not seem to fall out in a straightforward way from the coupled oscillator model of syllable structure. Having said that, one has to keep in mind that existing studies on differential onset-coda overlap in consonant sequences suffer from the for many languages unavoidable confound that consonant order reverses between onset and coda for the investigated language and hence the locus of this effect is not clear.

Secondly, while in some studies, the predictions of the coupled oscillator model of syllable structure could unequivocally be supported (e.g. Hermes et al., 2013), in other studies the

predictions were not fully confirmed. This has brought attention to a major weakness in the c-center effect measurement heuristics that have conventionally been employed which simply evaluate the stability of different lag values measured from kinematic data: Variability in the anchor point relative to which across-condition relative timing is evaluated can act as a severe confound. Shaw & Gafos (2015; Gafos et al., 2013; Shaw et al., 2011) addressed this problem by combining experimental measurements with simulations for Arabic and English data, two languages that on theoretical grounds are expected to show different syllable parses for prevocalic consonant sequence (C.CV for Arabic and .CCV for English). They argue based on their simulations that temporal lag measures can by themselves only be limitedly informative about underlying stability patterns and hence by extension, coordination relations. Rather, it is only by fitting the experimental data to simulated levels of anchor variability that the common c-center heuristic can be considered to give reliable results.

But beyond non-trivial methodological issues on how to relate kinematic observations to hypothesized underlying structure (see e.g., Mücke et al., 2020 for a recent discussion), there are also some deeper theoretical issues that have been raised. This concerns for instance a possible interaction between syllable dynamics and contrastive dynamics in a way which is currently not foreseen in the model. Secondly, as we would like to argue, the coupled oscillator model of syllable structure cannot be complete without a clearer definition of the consonant-vowel distinction in dynamic terms and without addressing the issue of sonority. We will now discuss these two points in turn.

3.2. The interaction of time and space

The original idea of Browman & Goldstein (2000) of phase relationships expressing syllable structure assumed that these phase relationships would hold across the board, independently of the particular segmental composition of a given syllable. Generally, in the hierarchy of the Articulatory Phonology model, coordination relations within a syllable are specified lexically by virtue of invariant phase relationships; there thus is no possibility for interaction in their model between structurally determined phase relationships and the particular tract variables that enter these phase relationships. Yet recent research for instance by Pastätter (2017; Pastätter & Pouplier 2017) , Shaw & Chen (2019) and Liu et al. (2022) suggests that interaction between mechanical and timing models need to be accounted for.

Pastätter (2017; Pastätter & Pouplier, 2017) was able to show based on Polish that the coarticulation resistance of an onset consonant may be a crucial factor in determining timing relations between onset and vowel. He argued that bonding strength – a weighting factor on phase relationships introduced by Browman & Goldstein (2000) but never further elaborated – modulates phase relationships depending on a consonant's coarticulation resistance. In his proposal, every consonant is specified for a coupling strength parameter that specifies how much dominance that consonant exerts in the case of conflicting phase specifications (as they occur in complex onsets). On the basis of articulatory modelling using the task dynamic synthesizer, he demonstrated that the relative strength of C1–V and C2–V coupling relations in a C1C2V onset is the determinant of the extent to which C2 (the vowel-adjacent consonant) increasingly overlaps with the vowel under increasing onset complexity.

Further probing into the question of whether intergestural timing varies as a function of spatial effects, Shaw & Chen (2019) presented a study on CV syllables in Mandarin. They found that if the vowel articulator happens to be in proximity to the target, vowel movement initiation was delayed relative to the consonant. They draw a possible connection to different rest positions of the tongue, modelled as neutral attractor, from which movement is initiated (for a discussion of the notion of neutral attractor in Articulatory Phonology, see Tilsen, 2019). Likewise they entertain the possibility of "downstream targets" (p.10f.), meaning that the phase relationships of gestural coordination may not necessarily refer to movement onsets but might refer to later targets – for instance the target achievement of the vowel may be timed to the release of the consonant instead of consonant and vowel movement being timed directly to each other (for the initial proposal of variation in landmark coordination, see Gafos, 2002, for a critical discussion of the question of onset vs. target timing see Turk & Shattuck Hufnagel, 2020). Another, highly interesting alternative they raise in passing (p.13) is the possibility of the timing of the vowel onset being determined by the distance to the target. They note that this is predicted by the nonlinear formulation of the gestural equation as proposed by Sorensen and Gafos (2016), underscoring that the question of how to cast the basic equation for a gesture has profound implications for all aspects of the model. However, Liu et al. (2022) could not replicate Shaw & Chen's (2019) finding: By applying a novel methodology for determining C-V timing across conditions, they are able to confirm global C-V synchrony for Mandarin Chinese. Yet they also make the point that C-V synchrony depends on the degree of spatial overlap (i.e., overlapping articulator recruitment) between consonant and vowel.

Let us illustrate this last point by way of an example. Generally, mechanical issues relating to how articulators achieve contrastive goals should not be a determiner of when a gesture *begins*. Yet this does not seem to be generally the case. To see this, we can compare two specific examples of a CV syllable /pa/ and /ta/, which should, according to the gestural syllable model, both show 0° phase relationships. For /pa/, lip closure for /p/ should thus occur around the same time as tongue backing for /a/. And indeed, Iskarous et al. (2010) showed that the horizontal motion of the tongue body for the vowel begins at the acoustic closure for the /p/ in /pa/. This is consistent with the 0°-phase model of C and V synchronization. But they also demonstrated that the horizontal motion of the tongue body for the vowel begins after completion of the /t/ in /ta/. The tongue first pushes the tongue tip towards the alveolar ridge for /t/, then starts to back up for the /a/. The /a/'s tongue backing can only begin when the /t/'s closure has been achieved. Therefore the /t/ and /a/ cannot begin together as predicted by the syllable model, as there is a mechanical incompatibility between the tongue 's fronting for /t/ and backing for /a/. This example serves to illustrate that spatial aspects may possibly constrain the temporal entrainment of gestures which currently the model does not allow for.

3.3. Phonological primitives and motor primitives – consonants, vowels, and 0° phase

Our second point relates to modelling some of the most basic properties of spoken language which a phonology has to account for: The gestural syllable theory, like any theory of phonology, builds on a fundamental distinction between consonants and vowels. Coordination relationships between gestures thereby crucially depend on this distinction. The gestural

molecule forming an onset, for instance, consists solely of consonants which are coordinated with at least one vowel. For postvocalic consonants, only the consonant immediately following the vowel is directly phased to that vowel; any further consonants are coordinated sequentially to each other. This means that the nature of the pairwise local coupling relations within a syllable strongly depend on the consonantal or vocalic identity of a given gesture. But, where does that distinction come from within a dynamical model without C and V being simply phonetic labels for classes of specific sets of gestures considered consonantal or vocalic within linguistic phonetics and phonology? Articulatory Phonology has made through its contrast dynamics a serious attempt to move away from thinking of linguistic contrastive specification as revolving around phonological features as classificatory labels, and rather model contrast based on spatio-temporal goals. That C and V are simply classificatory in nature, therefore, would not be consistent with that general approach to contrast. The alternative assumption, which has at least implicitly been used in the theory (e.g., Browman & Goldstein, 1992), is that vowel and consonants are emergent classes of sounds based on certain ranges of stiffness and constriction degree parameter values. Yet there is only a stipulative connection between certain ranges of stiffness and/or constriction degree parameter values and syllabic functional behavior, just as in non-spatiotemporal theories of phonology (Clements & Keyser, 1983). A somewhat related issue is one of the most studied phenomena in syllable phonology prior to Articulatory Phonology: sonority profiles and sonority laws of syllabic composition (e.g., Clements, 1990; Vennemann, 1988). There are plenty of exceptions to sonority-based generalizations, but the frequency of syllables in which sonority rises then potentially falls is too high across languages for a theory of phonological representations to not tackle this issue.[7]

Another issue we would like to raise relates to the assumed primitive status of 0° phase. Articulatory Phonology has had a strong motivation in relating typologically prevalent phonological phenomena to laws of motion and their perception, and this point can be seen quite clearly in a particular aspect of the coupled oscillator model of syllable structure: As mentioned earlier, one of the most prevalent typological generalizations regarding syllable structure is the universality of CV syllables. The hypothesized connection between language as motion and CV universality goes back to a set of rate-scaling experiments that have been interpreted to show that 0°-phase is the most stable temporal synchronization relation in human motor control, including speech (Stetson, 1951; Kelso et al., 1986). Since in syllable onset, consonant and the vowel start their movements (close to) simultaneously, there seems to be a

[7] In fact, in one of their earliest publications, Browman & Goldstein provide a very insightful discussion of this issue: In Section 3 of Browman & Goldstein (1989), the vocal tract is visualized in that section as a network of tubes in series and parallel arrangement. This is similar mathematically to a hydraulic, mechanical, or electrical network. In such a network, there are energetic flows constrained by series-parallel laws that determine, for instance, that the total flow in a series connection of tubes will be that through the narrowest tube, whereas the total flow in a parallel arrangement of tubes will be that through the widest. Due to these 'path of least resistance' laws, the vocal tract is no longer seen as composed only of a set of isolated, local constriction gestures. The arrangement of the constrictions/impedances and the laws of energy-conservation result in a global flow. Indeed, Browman & Goldstein (1989) argue that different consonants have different tube arrangements, and that their major class features like 'sonorant' and 'consonantal' could be due to the nature of the global flow due to each arrangement of tubes. These insights have yet to be integrated with the rest of the theory, but we would like to suggest that this proposal not only has the potential to lead to a principled treatment of acoustic and aerodynamic sources of influence on gestural coordination, it also leads to a possible origin of consonantal vs. vocalic oscillators, as much as a possible origin for the notion of sonority in the gestural framework.

natural connection to be drawn between a hypothesized 0° phase relationship between CV, its maximal stability and CV universality, establishing a direct link between motoric stability and typological preference.

Yet we believe that that the full range of human motor behaviors suggests that this connection may be too simplistic. Consider another very useful and more basic human behavior: locomotion. In that behavior, at slow speeds, our bipedal gait system almost always uses a 180°-phase relation between the legs, which is what we call walking. A more complex phasing relation we call running is used at high speeds, but it 's a variation on 180°-phase. What is almost never used in human locomotion is the 0°-phase gait: jumping, where the two legs move simultaneously. The same is true of quadrupedal gaits—there are many such gaits, with many different phase angles between the legs such as walking, trotting, pacing, and bounding (Golubitsky et al., 1999), but one of the rarest such gaits is the 0°-phase angle pronk, even though it is possible and sometimes used by Gazelles. Therefore, in locomotion, which is one of the most basic of all motor control skills, 0°-phase is exceedingly rare in evolution, therefore the idea that motor control tends to seek the 0°-phase as a highly stable coordination strategy may seem implausible. We do not mean to suggest here that 0°-phase is not present in human motor control, as we can see it, for instance, in grasping, where all fingers move together.[8] What we conclude from these examples, however, is that the phase angles required for a skill are driven by the task at hand: grasping requires 0°-phase because of force and accuracy requirements for grasping, whereas musical performance requires near arbitrary phase angles between the fingers, since the desired harmony, melody, and rhythm of the music requires it. We take these examples to mean that the abstract motor task drives the effectors to have particular phase relations. But it seems problematic to assume that tasks use specific effector phase relations *because* these phase relations are intrinsically stable. Onset-vowel coordination may very well be 0°-phase, yet we are questioning the reasoning for it, and hence the explanation for CV universality. Rather, we suggest that the temporal coordination strategy depends on the task, so that 0°-phase between onset and vowel would be explained by some reasons *intrinsic to the syllable production task*, rather than some notion of primitive motor stability.

4. The Elasticity of Time: Prosodic Planning Dynamics II

As we have discussed so far, in the gestural model lexical entries – so-called gestural structures – specify which gestures a given word or morpheme is composed of and their pairwise coupling relations. While each gesture and its relation to other gestures has fixed underlying parameter values, these values can be scaled as a function of speaking demands, speech rate, prosodic position and the like. The specific scaling the lexical parameters undergo during a given utterance planning results in the gestural score which is the input to the task dynamic system. The ideas of how this scaling of gestural parameters may come about have undergone major conceptual developments in the 21st century, one of them related to modeling the effect prosody on the temporal evolution of articulator kinematics.

[8] Kelso (1995: 99f.) in fact argues that natural systems will rather exist near but not in phase-locked states, since the latter are too stable and do not allow for enough flexibility.

Byrd & Saltzman (2003) in their position paper *The elastic phrase* advanced a major theoretical proposal for the dynamic modelling of prosody based on the so-called π-gesture, a time-warping gesture which directly impacts the time course of constriction gesture activation at prosodic boundaries (see also Byrd & Saltzman, 1998; Saltzman & Byrd, 2000). This model has subsequently, in a first sketch, been extended to a more general class of prosodic gestures, called μ-gestures (Saltzman et al., 2008), which serve to modulate either spatial or temporal properties of the gestural planning oscillators as a function of prosodic events. Importantly, spatial and temporal modulation have become separate planning entities in this model, although the concept of a spatial modulation gesture has not been elaborated any further to date. We will therefore mostly be focused on the π-gesture here.

Prosody overall serves to structure an utterance by marking boundaries and prominence (Beckman, 1993). It is thereby generally accepted that prosodic structure is hierarchically organized ranging from syllable-level constituency (onset, rhyme) over various levels to the intonational phrase. The scope of a prosodic event such as a boundary is typically larger than a single consonant or vowel but rather extends across segment and syllable boundaries, even though the articulatory signatures of prosody are evident in local constriction formation and release. Indeed, constriction formation and release of a segment may be affected asymmetrically by boundaries. Fine-grained changes to articulator kinematics at the subsyllabic and even subphonemic level as a function of prosody have become a major argument when adjudicating between different approaches to modelling the signatures of prosody in articulation. General overviews on this topic can be found, among others, in Fletcher (2010), and Turk & Shattuck-Hufnagel (1996).

4.1. The π-gesture model

A major point Browman & Goldstein made in their model of the syllable was that syllabic structure is emergent from the *pairwise local* phase relationships between gestures; there are no hierarchical association lines expressing the grouping of segments in the prosodic hierarchy. While this restriction to locality may capture syllabic effects effectively, higher-level prosodic events are known to be non-local in scope, and this has served as a major argument in favor of modelling prosody externally to constriction gestures by a separate prosodic gesture, the π-gesture (among others, Byrd & Choi, 2010; Byrd, et al., 2005; Byrd & Saltzman, 2003; Cho et al. 2017; Cho et al., 2014; Katsika, 2016; Katsika et al., 2014; Krivokapić, 2014; Lee et al, 2006; see Krivokapić, 2020 and Byrd & Krivokapić, 2021 for an overview).

The π-gesture, as proposed by Byrd & Saltzman (1998, 2003), provides a dynamical model of prosodic boundaries (as opposed to prominence). It enacts prosodic boundaries of all levels by inducing a slow-motion effect ('clock slowing') in any temporally coextensive constriction gestures. For instance, pre-boundary lengthening, which is a major correlate of phrase boundary marking (e.g., Cho et al., 2014; Katsika, 2016; Turk & Shattuck-Hufnagel, 2007) in this model is due to the π-gesture modulating the flow of time of any co-active constriction gestures (their 'internal clock') meaning the π-gesture slows down a particular stretch of the gestural score. This is schematically depicted in Fig 3.

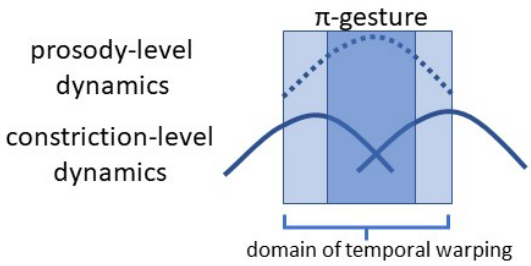

Figure 3. Schematic illustration of the π-gesture being concurrently active with the activation functions for two constriction gestures. The lighter shading of the rectangle represents domains during which the effects of the π-gesture will be relatively weak to its activation being ramped on and off. During the darker shaded area, the effect of the π-gesture on the time course of gestural activation of the constriction gestures is maximal. Observe how each individual constriction gestures is affected in a non-uniform fashion over its life-cycle by the π-gesture due to this ramped activation (see Section 2.1. on the concept of ramped activation).

This 'slow motion effect' only affects the time course of a given gesture's activation, not the strength of a given constriction gesture. The assumption is that these temporal changes in the activation trajectories affect kinematically both temporal and spatial changes in terms of maximal position and peak velocity – well-known spatial strengthening effects at boundaries as well as changes to gestural overlap at boundaries are in this model a by-product of the π-gesture induced time warping. Different boundary levels of the prosodic hierarchy (ip, IP) in this approach are a matter of π-gesture activation strength.

A recurrent issue in the context of the π-gesture is the question of its alignment. Byrd & Saltzman (2003) propose that the π-gesture should be assumed to be centered at the boundary, that is, centered on the constriction gestures adjacent to either side of the boundary. But this alignment, as they discuss, does not fall out from any property of the π-gesture itself, nor of the nature of the prosodic boundary; this is a stipulative working assumption made to constrain this overly powerful model. Burroni & Tilsen (in press) have recently argued for an asymmetric anchoring of the π-gesture in the case of stress-clash.

4.2. Prominence and boundary

Prominence, like boundary, induces systematic spatio-temporal variation in the speech signals which has aptly been described as localized hyperarticulation – articulators in stressed position reach more extreme positions compared to unstressed position (de Jong, 1995). Also the sonority-expansion description of prominence (Beckman et al., 1992) generalizes the observation that prosodic prominence systematically influences the spatio-temporal characteristics of local constriction gestures. The influential paper of Beckman et al. (1992; also Edwards et al., 1991) tested the viability of the task dynamic approach for modelling prominence and boundary. They came to the conclusion that the kinematics of accent are best understood in terms of a phase change between closing and opening cycles of a gesture instead of a local change to a gesture 's dynamic parameters. Their study found that accent affects both

the amplitude and the duration of both opening and closing gestures. Crucially, the opening, but not the closing gesture is faster in the accented syllable. They took this to mean that the opening gesture in an unaccented syllable is truncated by the closing gesture, whereas in an accented syllable this is not the case, giving the opening gesture more time and space to reach its underlying target.

Beckman, Edwards and colleagues further observed that there is a fundamental difference in the kinematics of accent and boundary. While accent affects the relative phasing of the closing and opening gestures to each other, the durational effects associated with boundary are not distributed equally across the syllable – the parts of the syllable closer to the boundary (i.e. the closing gesture in /ɑp#/) is affected more by the boundary-induced durational lengthening compared to the opening gesture. In contrast to accent, they find no significant spatial effect associated with boundary. They interpreted this as a boundary-induced change in stiffness of the closing gesture. While such a scaling of a gesture's natural frequency as a function of prosody maintains the principle that prosodic organization is expressed locally (as Browman & Goldstein proposed in the context of the syllable), the prosodic event itself is in this approach external to the dynamical system. As Byrd & Krivokapić (2021) point out, stiffness is in this view a mapping parameter of an extrinsic prosodic event onto local constriction kinematics. The prosodic event itself is not expressed in dynamic terms, and this was one motivation for Byrd & Saltzman's proposal of the π-gesture.

The other argument brought forward by Byrd & Saltzman (2003) against a stiffness implementation of prosody is an empirical one: A modulation of stiffness cannot account for a reduced amount of overlap seen between successive gestures which may be seen at boundaries (Bombien, 2011; Byrd & Choi, 2010; Cho et al., 2017; Guitard-Ivent, et al., 2021). To our understanding, the π-gesture accounts for by this phenomenon by virtue of its ramped activation: The closer a given (part of a) gesture is to the maximum strength of the π-gesture, the more that part will be slowed down.[9] Partial overlap of a gesture with the π-gesture will thus lead to a nonuniform lengthening effect over the course of that gesture. Hence for instance constriction formation, but not the release of a post-boundary consonant may be slowed down, resulting in a reduced degree of overlap. Yet the effects of boundary on overlap have been quite inconsistent across speakers, as Guitard-Ivent et al. (2021) point out. In their study, which comprises a large corpus of acoustic data, they are able to confirm for French /a/ that prosody modulates coarticulatory overlap, but they propose this to be dependent on the specific coupling relations within a syllable which may act as a resistor against prosodic modulation. Here again, like in our section on syllable structure, we are confronted with the possibility that there is feedback between the constriction-level and the prosodic level instead of strictly feedforward control.

Another point that is important to note here that when Beckman et al. talk about accent modulating inter-gestural phasing, they refer to the opening and closing gesture of a single consonant or vowel (for such a separation of constriction formation and release into separate

[9] Note that this account relies on non-autonomy as defined earlier in Section 2. If the theory is to be autonomous in the gestures, contrastive, and prosodic, a nonlinear theory of the prosodic gesture, where the natural frequency varies based on the within-gestural state, rather than time, would need to be proposed.

control units, see the split-gesture dynamics of Nam et al., 2007), whereas Byrd & Saltzman (2003) assume gestures to be a single control unit, i.e. constriction formation and release for a given consonant or vowel follows from a single gesture with a uniform dynamic parameter specification. In the Byrd & Saltzman model the concept of truncation by overlap cannot be applied meaningfully within a single consonant or vowel (even though constriction formation and release parts of a gesture may be affected to by the prosodic scaling to a different degree due to the π-gesture's ramped activation and depending on its precise alignment). The cross-study discrepancy in description levels of overlap is more than a glitch: It serves as a poignant reminder that the questions of how to define a constriction gesture and of how to model the signatures of prosody in articulation can ultimately not be solved separately, even though studies usually deal with one topic or the other, since any given study can only deal with a certain level of complexity.

Roessig & Mücke (2019) have further advanced our understanding of different types of prominence within a dynamical system. They investigated tonal and constriction-level modifications to utterances in German as a function of unaccented vs. accented and, within the accented condition, three types of focus (broad, narrow, contrastive). In their results the distribution of f0 contours changes from flat (unaccented) to bimodal in the accented, broad focus condition which can have either a rising or falling pitch accent. Within the accented conditions, the distribution changes further as a function of focus type with narrow and contrastive focus having an increasing dominance of rising f0 accents. At the constriction level, this tonal variation goes hand in hand with increasing lip aperture and a lowered tongue body position for the (low) vowel. The contribution of this study is to provide a unified mathematical description of these results in both domains (tonal, constriction level) within a single dynamical system: Under continuous change of a single control parameter (k) which enacts the varying focus conditions, the probability functions of both experimental variables – tonal pattern and lip aperture – are affected, implementing mathematically the observation that the impact of prosody is multi-dimensional: it is signaled in pitch contours as much as in changes to the articulator kinematics. Focus categories in this view correspond to stable attractor basins in a dynamic landscape which arise as a function of variation in a continuous scaling parameter that simultaneously, in this study, enacts accent and focus. This work demonstrates elegantly how variation in continuous phonetic parameters which are usually treated independently of each other (laryngeal vs supralaryngeal kinematics or pitch vs constriction degree) can be obtained from scaling a single 'prosody' parameter (see also Gafos 2006, Gafos & Benuš, 2006). The categories themselves correspond to the ones established in the prosodic hierarchy, and can be understood within the language of dynamical systems.

4.3. μ-gestures

The π-gesture provides a model of boundary, a similarly elaborated theoretical proposal for how to modulate accent in dynamic terms is, besides the proposal by Beckman et al. mentioned above, missing to date. In a first proof-of-concept, Saltzman and colleagues (Saltzman et al., 2008), building on a proposal by O'Dell & Nieminen (1999, 2009), recast the prosodic hierarchy in the language of dynamical systems as a set of hierarchically nested oscillators (see also Barbosa, 2002). They also generalized the idea of the π-gesture to a general class of

modulation gestures, termed μ-gestures, (of which the π-gesture is seen to be one particular kind) that either modulate time or spatial parameters. These modulation gestures are designed provide a model of both prominence and boundary in dynamical terms. The μ gesture idea has hardly been developed, but since it is in current literature being used as tool to understand interactions between prominence and boundary, we include it in our review here.

The μ-gesture is designed to ultimately be a generalization of the π-gesture to a class of temporal modulation gestures. Temporal differences between stressed and unstressed syllables are in this view the result of a temporal modulation μ-gesture being active during a stressed syllable. As is the case with all classes of gestures, the activation functions are ramped on and off, meaning the strength of the μ-gesture increases and decreases gradually at the edges of the activation interval. This μ-gesture operates at the level of gestural planning and modulates the stiffness of a syllable-planning oscillator, leading to a local change in phasing between syllable-foot oscillator ensemble. Note that in this model, phonological units of the prosodic hierarchy (syllable, foot) have their own planning level oscillators that are designed to trigger (in ways not specified yet) the constriction gesture activations. They also allow for interactions between these different prosodic levels, although currently not with the constriction level dynamics – a point we have already critically remarked on earlier. This means that for instance, a μ-gesture active for the stressed syllable within a foot modulates the natural frequency parameters of both the foot and the syllable planning oscillators – this is in contrast to the π-gesture which operates directly at the level of the gestural score and acts to lengthen all concurrently active constriction gestures.

The Saltzman et al. (2008) contribution to modelling prosodic structure is a proof-of-concept for the modelling of the traditional prosodic hierarchy as hierarchical ensemble of oscillators. The prosodic hierarchy is in this approach an collection of time-keepers that interact with each other (such as the foot and syllable-level oscillators) and that ultimately trigger the constriction gesture dynamics. Saltzman et al. also hint at the possibility of a *spatial* μ-gesture that might serve to warp spatial target parameters of constriction gestures. The theoretical concept of a class of μ-gestures has to date not been further elaborated, even though the idea of the temporal μ-gesture has found recent application among others in the work of Katsika (2016).

How interactions between the distinct types of prosodic gestures posited by the Byrd & Saltzman team may be envisioned to impact articulator kinematics has been subject of a study by Katsika (2016). She provided evidence that the scope of pre-boundary lengthening varies with the location of the lexically stressed syllable: in her data on Greek, pre-boundary lengthening started earlier in words with non-final stress compared to words with final stress. She also observed that constriction formation and release of a given consonant or vowel may be asymmetrically affected by the boundary, just as Beckman et al. (2009) had observed (spatial effects are not investigated in Katsika's study). She thus proposed for the π-gesture to be coordinated with both the phrase-final vowel as well as with the μ-gesture of the last lexically stressed vowel preceding the boundary. If this stressed vowel is not the last vowel of the phrase, this coupling will cause the μ- and the π-gestures to shift towards each other. By virtue of each of these prosodic gestures also being coupled strongly to the respective vowels, the shift is only subtle: pre-boundary lengthening occurs earlier, and near or on the μ-gesture a shortening effect

is observed relative to a condition where π-gesture and μ-gesture coincide (when the stressed vowel is the last vowel before the boundary). Katsika thus uses the concept of the μ-gesture to describe how the π-gesture may vary in its alignment between utterances that have the same type (or degree) of prosodic boundary but differ in the prosodic structure of the material preceding the boundary – crucially, in the distance of the last stressed syllable to the prosodic boundary (for a proposal of asymmetric π-gesture alignment without assumption of a μ-gesture, see Burroni & Tilsen, in press).

While this certainly is an appealing idea, there is at present no computational model that would allow us to test these hypotheses and make quantitative predictions. Also note that the μ-gesture and the π-gesture in their current inceptions do not operate at the same level (see Byrd & Krivokapić, 2021 for discussion): the π-gesture directly affects the gestural score which is the result of gestural planning, while the class μ-gestures are part of the hierarchical ensemble of prosodic planning oscillators. How a modulation gesture that affects the natural frequency of utterance planning oscillators would be able to affect the alignment of the π-gesture has not been addressed yet.

Byrd & Saltzman's dynamical model of boundary has been impactful and has contributed to our principled understanding of the signatures of prosody in articulation, yet currently the modelling of the grouping of segments in the prosodic hierarchy, prominence, and boundary has proceeded largely independently of the constriction dynamics and of the coupled oscillator model of syllable structure. While the concepts introduced with the π-gesture have opened an entirely new research avenue for a dynamic model of prosody, it is in its current form an overly powerful model with too many degrees of freedom. The vision sketched in Saltzman et al. (2008) of separate spatial and temporal μ-gestures and the ensuing categorical separation of the control of spatial and temporal planning parameters will require careful motivation and extensive computational modelling work to turn this into a model with sufficient predictive power. In particular the impact these complex ensembles of planning oscillators might actually have on the tract variables of constriction location and degree remains to be worked out. Byrd & Saltzman (2003) have contributed tremendous insights on how to incorporate the prosodically conditioned elasticity of time into Articulatory Phonology, yet many aspects in the dynamic modelling of prosodic structure remain to be worked out.[10]

[10] In a recent critique of how prosodic timing is accounted for in Articulatory Phonology and Task Dynamics, Turk & Shattuck-Hufnagel (2020, 2021) point to non-speech motor control skills (Lee, 2011) whose planning is presumably based on their temporal endpoint, instead of onset of movement. Specifically, timing variation is greater towards the beginning of an action than towards the planned ending (goal). Potentially, this could be a fruitful approach for understanding the "why" of final lengthening, and the dynamical differences between beginnings and endings of actions. This is especially the case since the work pointed to by Turk & Shattuck-Hufnagel (2020), Lee's (2011) general tau-theory, is framed in the general dynamical approach to motor control pointed to earlier, with its roots in the Bernsteinian approach to motor control and the Gibsonian approach to direct perception. But unlike Turk and Shattuck-Hufnagel (2020), we believe that attention to the endpoint triggering idea from general tau-theory would not lead to evidence for a phonology-extrinsic timing mechanism. This is because the Articulatory Phonology spatiotemporal approach takes speech to be *phonological action*, and that such linguistic actions could have properties of other non-speech actions.

5. Planning Time

Evidence has accumulated over the decades that words are planned in that they are assembled from smaller components (see Goldrick 2014 for an overview). A dynamical theory of the nature of the planning process has been developed over the last two decades within Articulator Phonology (Saltzman & Munhall, 1989; Nam & Saltzman, 2003; Goldstein et al., 2007; Tilsen, 2016, 2019; Roon & Gafos, 2016) and also within the wider framework of dynamical approaches to cognition (Grossberg, 1973, 1978). Several mathematical approaches have arisen within Articulatory Phonology to account for different aspects of this planning process, but they all share a core commitment with Fowler et al. (1980) and Fowler (1985), the first explorations of what a theory of planning within a non-dualistic phonological theory of speech production and perception should look like. Fowler argued that performance and planning must be parts of a *united* non-dualistic process, as opposed to processes of different natures with an interface of translation between two sets of laws. Her main argument was that phonological representation and computation tacitly but deeply refer to many facts about the working of the performance systems, and that the latter work the way they do due to their carrying out a linguistic task. The two are organically and inextricably linked.

Speech planning needs to account for the nature of phonological representations as much as for their retrieval and serial ordering. For the former, speech errors have served as an important source of evidence, and have been taken to confirm the psychological reality of linguistics units (among many others, Shattuck-Hufnagel, 1979). Speech errors have long been analysed as errors in serial ordering of symbolic segments during phonological planning; an overview is provided by Meyer (1992). Speech errors are thus often seen as a source of evidence for cognitive representations of phonemic knowledge requiring symbolic representations, and against dynamically defined units being units of utterance planning. Goldstein et al. (2007) were able to show for the first time that at least some speech errors are compatible with the view of dynamic gestural planning units: Detailed observations of articulatory and acoustic records of speech errors suggest that what may sound or be transcribed as a symbolic segmental replacement at least in some cases arises from a simultaneous co-existence of multiple planning units in the vocal tract (see also Pouplier, 2007, 2008). This was interpreted against the background of basic stability patterns known from general coordination principles and synchronization research (e.g. Pikovsky, et al. 2001): complex movement and coordination patterns may break down (for instance under increasing rate) and revert to a 1:1 frequency, in-phase pattern. This, as was argued by Goldstein et al. (2007), can be observed in speech errors under the assumption that planning oscillators (Saltzman et al,. 2006) in error-inducing environments come to be attracted to this natural stability pattern, resulting in a simultaneous production of gestures which would otherwise be sequential (see Section 3.1 on the assumption that basic stability patterns in nature are relevant for speech). However, this account of speech errors based on general dynamical principles has to date not been integrated any further into a gestural model of phonological planning.

As to serial order, utterance planning requires vocal tract gestures to be *coordinated* or organized with respect to each other temporally. The initial theory of Articulatory Phonology (Browman & Goldstein, 1986, 1989) had assumed a given gestural score, specifying when each

gesture ought to begin and for how long it should last, stipulating the coordination relations. As early as 1989, it was felt that the timing of gestures ought to arise as part of a dynamical process of interaction amongst the gestures, rather than being externally stipulated. The linguistic system will still need to specify the linear order of contrasts, distinguishing, for instance, the utterances /sap/ from /spa/ or /psa/, but this overall order leaves a great deal underspecified regarding the timing of the gestures and their level of overlap, and it is the computation of this coordinative timing that a theory of planning needs to accomplish.

The section titled "Serial Dynamics" in Saltzman and Munhall's (1989) original paper on Task Dynamics was an initial attempt at elucidating how such a coordination process could work, inspired by the dynamical work of Grossberg on intrinsic serial ordering (Grossberg, 1978, 1987) and Jordan's recurrent neural networks (Jordan, 1986). In the words of the authors: "Explaining how a movement sequence is generated in a connectionist computational network becomes primarily a matter of explaining the patterning of activity over time among the network 's processing elements or nodes. This patterning occurs through cooperative and competitive interactions among the nodes themselves." (Saltzman, & Munhall, 1989 p. 356).

This initial outline was replaced by a very explicit theory of timing coordination in Browman & Goldstein (2000) and Nam & Saltzman (2003) in the form of the planning oscillator model of syllable structure, as we have discussed earlier. The idea of "cooperative and competitive interactions" initiated by Saltzman & Munhall (1989) has become part of the theory of planning as entrainment amongst limit cycle oscillators in a coordinative -competitive process. Dynamic Field Theory (DFT) (Erlhagen & Schöner, 2004), which also has a coordinative-competitive logic, has been used by Roon & Gafos (2016) to model how perception can influence production planning as well as by Tilsen (2016, 2019) to model serial order and selection. We will now provide a synopsis of each of these 21st century theories of planning.

In the initial exposition of Articulatory Phonology, the gestural parameters, task target and stiffness, were simply provided as static, phonologically given parameters. Experimental results on perceptuo-motor effects, however, seem inconsistent with such a model, as what a speaker perceives as they speak may affect online the fine phonetic details of speech production (e.g., Yuen et al. 2010). Galantucci et al. (2009) designed an experiment in which the participants were taught a syllable to say when a particular visual stimulus (e.g., ##) is presented, establishing the intended syllable. But as the participant was about to speak, they heard a distractor very shortly before they respond. The basic result was that the reaction time was long when the distractor was incongruent with the intended response, and short when the distractor and response were congruent. Roon & Gafos (2015) extended the results beyond identity to show that even if a contrastive unit, Voice or Vocal Tract Articulator, is shared between distractor and response, the reaction time is shorter. Since the lag between the distractor and response is so short, this is taken to mean that the perceived and intended utterances interact, somehow, to yield the actually produced utterance. That is, a gesture 's parameter value during performance may actually be different from the linguistically-intended one, suggesting that some dynamical process occurs that determines the actually produced parameter value based on some competition between the linguistically-intended parameter value and some other perceived value. Roon & Gafos (2016) provided a model for this

interactive planning process, using DFT (Erlhagen and Schöner, 2004), a dynamical theory of action, perception, and cognition. This theory grew in the same dynamical milieu (Grossberg, 1973, 1978; Turvey, 1990) as Articulatory Phonology, and currently is an important component of understanding planning within Articulatory Phonology, as it has been extended further by Tilsen (2019). Therefore, we will also give a synopsis of DFT. We refer the reader to Schöner et al. (2016) for a detailed exposition of DFT, and to Roon & Gafos (2016) and Tilsen (2019) for details on its extension to speech.

At the foundation of DFT is the presence of two fundamental variables, a metric variable and an activation variable. The metric variable could be a movement parameter like a direction of motion or a percept like a color or sound quality, and each such variable is defined on a metric scale of possible movement or perceptual qualities. When applied to Articulatory Phonology, the metric variables are the differential equation constant coefficients, which at implementation time are constant, but can change in the dynamical process of planning. The metric scale is then all the possible values for LA_0, for instance. The activation variable specifies the degree of salience of each value of a metric variable. The latter variable is dynamical, and the planning process determines how the salience or activation of each metric value changes through coordination and competition. At the end of planning, the metric value with the highest activation is chosen as the task parameter value for performance. The DFT model of the planning process is in full conversation with the performance process, as called for by Fowler (1985). Its non-dualism is evident in the tight interweaving of the independent variable (metric) with the dependent variable (activation) as elements of one function.

Before the planning process, activations for all possible values of each metric variable are similar and below a threshold required for implementation. At the initial time of planning, the linguistic system would provide an activation value above threshold at the intended metric value. For instance, if the linguistic system wants to specify that a /p/ is uttered, then there would be a peak activation at the LA_0 intended metric value for that segment. But there could be other peaks of activation that for instance the perceptual system could provide. If the distractor's metric value is congruent to the response, then the peaks of activation will coincide, whereas if it is incongruent, the activation distribution will be bimodal. There are four main aspects of the coordinative-competitive dynamical system whose input is the initial activation, and whose solution is the final activation profile. First, like fire, activation peaks promote their own growth and spread, an effect called *local excitation*. Second, metric values far away from each other inhibit each other by an amount related to their activation. This competitive aspect of the dynamics, called global inhibition, leads to peaks being able to inhibit other peaks. Third, a sigmoid is used to specify an activation threshold below which activation values are too low to play a role in the dynamics, leading to their suppression. Fourth, each metric value or task in a complex task suppresses other tasks, serving to promote its own salience. Fifth, there is a Monitor that will trigger the performance system with a metric value if its activation value hits a particularly high value. These dynamical factors are implemented mathematically as a differential equation for the evolution of each activation variable at each metric value with respect to time, due to the interactions just listed.

The explanation of the Galantucci et al. (2009) results in terms of DFT is as follows. When the participant sees a visual cue, their linguistic system initiates planning with a subthreshold activation peak at the intended value of the metric variable. Local excitation would make that peak get higher and higher and potentially trigger the Monitor, but as the peak gets closer to the Monitor-specified value for triggering implementation, an incongruent distractor provides another peak activation at a different metric value. Global inhibition then leads to competition between these two peaks, with the incongruent peak delaying the rise of the intended peak, even if the intended peak is higher. And it is this delay that leads to a longer reaction time in the perceptuomotor experiment. Congruent distractors would not delay triggering of performance. Therefore, DFT provides a dynamical explanation of the speech planning process, allowing for perceptual feedback to compete with linguistic intentions. The mutual inhibition amongst simultaneously active tasks, as in the Fourth dynamical factor above, is used by Roon & Gafos (2016) to explain the results of Roon & Gafos (2015) on how congruent distractors need not share all contrastive values with the response. State Feedback Control and Reinforcement Learning (Ramanarayanan et al., 2016; Parrell, 2021) have also been introduced to account for speech production perturbations due to acoustic feedback (Houde & Jordan, 1998). These are important models of online modification of performance, but for space-reasons, we will not discuss them further. It will be interesting to see if DFT and these models could merge, as planning and performance use the same fundamental dynamical language within Articulatory Phonology.

Tilsen (2016) used Grossberg's interactive theory of action selection and action serial ordering (Grossberg, 1978; Houghton, 1990) to generate a general theory of task interaction during planning in Articulatory Phonology, and Tilsen (2019) has extended this further using DFT to an explanation of a variety of local and long-distance phonological phenomena. Grossberg (1978) and Fowler et al. (1980) revived Lashley (1951)'s critique of the chaining approach to serial order, and argued for the need for a theory in which subactions that are fluently serially ordered with respect to each other to yield a larger action all interact amongst each other. These interactions, both coordinative and competitive, should yield the particular way in which the larger action is composed of the subactions. Grossberg (1978) developed an interactive neural theory in which each subaction is represented by a mathematical neuron or node, which can be active to varying degrees. The magnitude of the initial activations of the subactions leads to the order of the subactions. Specifically, *higher* activation of a node leads to *earlier* activation within the plan, and greater control over the performance system.

As we have discussed for DFT, there is a threshold logic, which governs how high an activation needs to be to actually trigger performance. Also, as in DFT, there are local excitatory coordinative and global inhibitory competitive forces, where currently active nodes promote their own growth and inhibit the other nodes. Grossberg (1978) discusses the many ways in which noise and inputs to this interactive dynamic between subactions yield to different aberrations in planning, and Tilsen (2016, 2019) shows how local and non-local assimilation phenomena in phonology, the typology of alternation, and various generalizations regarding the development of coarticulation follow from these dynamical interactions among subactions, here interpreted as gestures.

From the initial outline of Saltzman and Munhall (1989) on how coordination and competition can lead to coordination, the later theories of Roon & Gafos (2016) and Tilsen (2016, 2019) have built very specific theories of the nature of these interactions. And this work has relied on the earlier work of Grossberg (1978) and Fowler et al. (1980) viewing planning not as a box and arrow process, but as a continuous activation, interactive phenomenon, a view that has matured into DFT.

6. Acoustic Phonology

"Speech as Audible Gestures" is a dictum of Stetson's (cited by Löfqvist, 1990) that has been at the basis of Articulatory Phonology's attempt to answer the central questions of Linguistic Phonetics mentioned at the outset through the models of how language structures the spatial and temporal characteristics of *articulation*. But we believe that for a theory to truly capture the first principles of linguistic phonetics, there needs to be a full embracement of the audibility of gestures, the acoustic and perceptual sides of the skill of speech. In this final section we argue that an Acoustic Phonology is as necessary as an Articulatory Phonology for characterizing that skill.

Articulatory Phonology grew within a general ecological approach to speech communication that includes a theory of perception – Carol Fowler's *Direct Realism* (Fowler, 2018; Fowler, 1986; Best et al., 2016), balancing Articulatory Phonology's concentration on articulation. Goldstein & Fowler (2003) presented in what they termed *Public Phonology*, a combined a theory of articulatory action and direct perception. In direct perception, animal perceptual abilities are seen as especially attuned to the objects and events in the world they perceive, which in this case are linguistic articulatory objects that cause the percepts. Direct perception received its initial support from psychological experiments which attempt to solve puzzles in human speech perception by arguing for perceptual recourse to the articulatory gestures that generated the percept (Liberman and Mattingly, 1985; Fowler, 1986). In the 21st century there have been developments in other areas of science, neuroscience, machine learning, and speech development suggesting that theories of perception centered on the articulatory system are necessary for understanding speech communication.

In neuroscience, there are results such as those of Assaneo & Poeppel (2018) who showed that processing in auditory and motor lobes are highly synchronized, suggesting a deep computational *parity* between production and perception (Liberman & Whalen, 2000). This work also finds evidence for the phonological syllable, a central organizer of articulation in Articulatory Phonology, as a major locus of connection between articulation and perception. In machine learning, the achievement of higher recognition accuracy, robustness to noise, and the achievement of systems requiring far less data than current ones, in vision, speech, and language technologies, has led to the recent development of learning algorithms that seek the *causal* mechanisms that generated the data (Pearl, 2019; Schölkopf, 2021). Since articulatory gestures in the vocal tract are the immediate causes for the acoustic signal, and since human perception, like automatic speech recognition needs to be robust to variation and likely to have evolved to be efficient, causal models of perception in human perception may be necessary as well. Work on development of speech in infants has also suggested deep engagement of articulatory structures in the perceptual process (Bruderer et al., 2015), and that some of the

earliest representations are multimodal (Choi et al., 2021). Work in speech signal processing suggests that articulatory and acoustic representations relate to each other in far more organic ways than previously thought (Yehia et al., 2002). Iskarous (2010), building on these notions, showed theoretically and empirically how formant frequency locations and amplitudes capture the same information as the anti-symmetric and symmetric components of the area function. However, despite Articulatory Phonology 's attempt to link speech production and perception, we believe that the theory has yet to connect these two aspects of spoken language communication sufficiently well to provide a truly Public Phonology. The reason for this, we believe, is that the connections between production and perception, aerodynamics and acoustics, the communicative output of production and input to perception, have not been given enough attention yet within Articulatory Phonology.

After all, the acoustic input to the perceptual system is a reflection of the global shape of the vocal tract (Fant, 1960), whereas Articulatory Phonology, as pointed out by Mattingly (1990), is overly focused on individual localized events. Some work has been done quite some time ago trying to define what an aerodynamic task would require (McGowan & Saltzman, 1995). We believe that renewed attention to these additional sources of influence on gestural timing and broadening the view of gestural goals will help our understanding of the skill of speech communication, involving language, production, and perception in an integrated whole.

If the direct realist approach to speech perception is fruitful, then changes in the acoustic signal ought to point to gestural changes. What then is the relation between the dynamical changes in currently active task variables and the dynamics of parameters of the acoustic signal like formants and spectral moments? There are indeed efforts in this direction already (e.g., Chartier et al., 2018), but we believe that far more is necessary. If sophisticated forward and inverse models could be built relating articulation and acoustics across a large array of speech contexts and languages, then a great deal of the articulatory process including coefficients like stiffness, target values, and phase angles could be discernable from the signal, allowing for the study of articulation through much more widely available acoustic signals. Turk & Shattuck-Hufnagel (2020) point to many works that have argued for acoustic, rather than articulatory, targets for at least some segments in some languages. And Iskarous & Kavitskaya (2018) argued for the role of audibility of gestures from the speech signal as an important constraint on sound change. It is only when the dynamical relations between articulation, acoustics, and perception have been far more fully worked out that worked out that we can understand the nature of targets and how articulation and acoustics interact with diachronic and synchronic phonology. Therefore, we believe that the construction of such models would yield a far-clearer theoretical understanding of the aspects of speech, and would also lead to a richer empirical basis for Public Phonology.

Concluding Remarks

The proponents of Articulatory Phonology and Task Dynamics contributed in the 1980s a ground-breaking conceptual innovation on how to think about spoken language as a phenomenon in nature, and the relationship between the cognitive and physical aspects of spoken language. Their framing of the skill of speaking within a dynamical systems approach

has had a deep influence on the field of phonetics, linguistics, and beyond. In this paper, we have reviewed some of the major advances in this theoretical approach from the past two decades, and have discussed some longstanding and some more recent open issues, relating to the distinction between consonants and vowels, the prosodic organization of speech, and speech planning. The rich, multi-faceted research that the gestural framework has inspired will undoubtedly continue to make a lasting impression on the field.

Acknowledgements. We are particularly grateful to Dani Byrd, Louis Goldstein, and Jelena Krivokapić for many insightful discussions on the topics addressed in this paper. All errors remain our own.
Competing interests. The authors declare no competing interests.

Appendix A: Dynamical Systems and Task Equation

A differential equation or a dynamical system describes how a system will evolve over time based on its current state. Differential equations are mathematical statements of *laws* of change. When they are *solved* we obtain a mathematical function describing how some phenomenon in the world unfolds over time. For instance, a differential equation describing the law that describes how an hour glass works could be stated verbally as: the change in the amount of sand in the upper chamber of an hour glass with time is some constant -*k*, let us say -10. This statement means that in every unit of time, 10 units of sand fall into the lower chamber. The negative corresponds to the change being a decrease. If x is the amount of sand in that upper chamber, we would symbolically express this law of the hour glass as $x_t = -10$, where x_t means the change in x with time. A phenomenon abiding by this law, a solution of the corresponding differential equation, could be described by a mathematical function with time as the independent variable, the amount of sand in the upper chamber as the dependent variable, and the function would be a line with a slope of -10. The y-intercept of this function is determined by how much sand is in the upper chamber before the first sand grain falls. If there were 1000 units of sand in the upper chamber, a solution function describing the phenomenon would be $x(t) = -10t + 1000$. At the first instant, when $t = 0$, there will be 1000 sand units (we call x at $t = 0$ the *initial condition*), at the second instant, when $t = 1$, $x = 990$, etc. So, we already see one full system with a law, $x_t = -10$, and a phenomenon in nature abiding by that law, $x(t) = -10t + 1000$. Iskarous (2017) provides a simple arithmetical way of obtaining this solution from the differential equation.

A different clock *contrasting* with this one in being faster, having a larger magnitude k, would, for instance, have the law $x_t = -100$, and its solution, for the same initial condition, would be $x(t) = -100t + 1000$. In this example as in other more complex physical laws, the differential equation law is about a change in a quantity (x in this example). A differential equation will usually have *constant* numbers, or parameters, like the k in this example. But when we solve the equation to obtain the function describing a phenomenon in the world, that function is a continuously varying function. So, in dynamical systems descriptions, the law has discrete unchanging parameters, but the solution function can be continuously variable. The reason that this is all relevant to speech is that phonological contrast constancy for the duration of a contrastive unit can be modeled in a discrete constant number like the k in a differential

equation, whereas the phonetically changing quantity, articulatory or acoustic, can be represented by a time-varying function like x, here.

Consider now a different dynamical law, the one describing the viscoelastic flesh on your forearm. If you leave it alone, it will stay there. If you push it, and remove your finger, it will return to its original configuration. If you pull it, and release, it will go back to its original position. This system has a preferred state and returns to that state when perturbed. This is a simple example of an equilibrium-seeking system, and there are many such systems in nature. Let us say that x is the position of the flesh, $x = 0$ is the resting position of the flesh, pulled flesh leads $x > 0$, while pushed flesh refers to $x < 0$. A simplified form for a differential equation describing such a system would be $x_t = -x$. If $x = 0$, then $x_t = 0$, and no change occurs. If we pull the flesh, $x > 0$, so $-x < 0$, so x_t is negative and x decreases, as expected physically. If we push, $x < 0$, $-x > 0$, and x_t is positive, so x increases. Loose material returns slowly, while tough material returns fast, so we say that each kind of material has a different stiffness k, and we amend the differential equation to $x_t = -kx$. If we are treating a system in which the equilibrium value x is not equal to 0 but some other constant value x_0, then the differential equation would be: $x_t = -k(x-x_0)$. Whenever x is different from x_0, this differential Equation says that there will be a force returning x towards x_0. Abstractly, there are two players in this description of nature: the constant coefficients x_0 and k, on the one hand, and the time varying x in the solution, which we have not written out explicitly here, but Figure A1 gives example solutions for different initial conditions for $k = .33$ (a), and $k = .8$ (b), for $x_0 = 30$ in both cases.

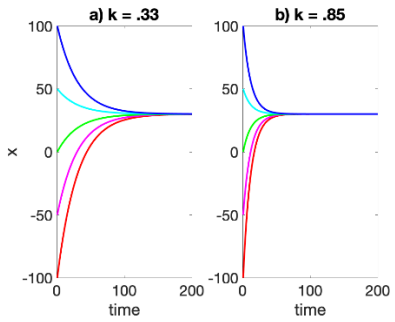

Figure A1. Equilibrium-seeking system solutions with equilibrium value $x_0 = 30$, and (a) $k = .33$ vs. (b) $k = .85$.

We now proceed from these simple examples to a simplified version of Task Dynamics that predicts the articulatory change in the following linguistic items: /ap/, /av/, /ip/, /up/, /aw/, /au/, as examples of how the theory works. Task Dynamical differential equations govern the change of a time-varying quantity called the *task variable*, usually a constriction location (CL) or a constriction degree (CD) defined somewhere along the vocal tract. For instance, the distance between the lips or Lip Aperture (LA) is a 'phonetic' time-varying task variable that is a CD defined at the lips. There are many other task variables for different places and manners of articulation (Browman & Goldstein, 1989). The constants in the differential equation for the task variable contain contrastive information provided by the linguistic system for some specified amount of time, the duration that the linguistic contrast is in control of the vocal

tract[11], and the linguistic differences amongst, for instance, labial stops, fricatives, approximants, and round vowels would need to be encoded in these constants. To accomplish a /p/ for instance, we want the lips to be closed, i.e. LA to become *0*[12], the *target* for /p/. But to accomplish a /v/, the target is a little greater than *0*. Now to make a /p/ or a /v/ after some vowel, then LA needs to decrease from its value during the vowel and approach the target for the consonant: if LA is far from its target value, it will change until LA achieved its target value. Therefore, achieving a contrastive target seems to be an equilibrium-seeking situation. A very simple differential equation that would achieve this is $LA_t = -k(LA - LA_0)$. This says that LA will decrease due to two factors: 1) the difference between the current LA state and the target LA_0, and 2) some constant k, called the stiffness, with larger k leading to faster decrease in LA. We will now explore the predictions of this dynamical model to see how its different parts are required for linguistic use of the vocal tract.

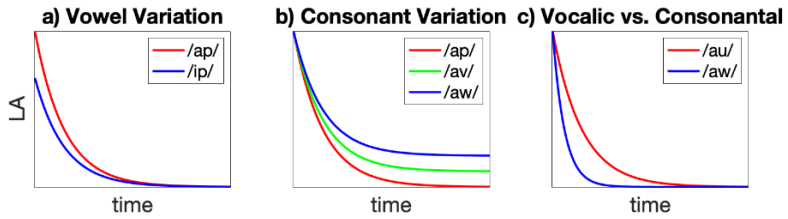

Figure A2. Linguistic effects of Initial Condition (a), targets (b), and stiffness (c).

Figure A2.a shows the predicted solutions for /p/ in /ap/ achieved from an initial conditions for /a/ (red) vs. /i/ (blue). /a/ requires a large LA therefore the distance between LA and LA_0 is initially large, while /i/ requires a smaller initial LA. Since /p/ requires one LA_0 in both /ap/ and /ip/, the only difference due to linguistic prior context is the initial condition. Note how dynamical concepts fit nicely with the linguistic situation. Figure A2.b. shows the predicted solutions for /ap/ (red) vs. /av/ (green) vs. /aw/ (blue). Each of the consonants here has a different LA_0, so even though they all start from the same LA for /a/, they go to different target values. An important difference between vowels and consonants is that the latter are achieved more quickly than the former. This can be achieved using a difference in k, with consonantal corresponding to higher k and vocalic to lower k. Figure A2.c. shows the achievements of /aw/ (magenta) vs. /au/ (cyan), where we have assumed for simplicity that the target values are not different, but only the k 's. We hope the reader sees again how more and more properties of phonology-phonetics correspond to basic equilibrium-seeking systems, allowing for a task dynamic treatment.

How well, though, do the predicted LA trajectories from the simplified task equation $LA_t = -k(LA - LA_0)$ fit actual articulatory data? The change in LA from /a/ to /p/ for an American English speaker can be seen in Figure A3.a. (top), and the actual velocity of change in A3.a. (bottom). And Figure A3.b. shows the LA prediction of the simplified task equation. As can be seen there is a qualitative difference in velocities. In the real transition, LA slowly increases then achieves a peak in velocity magnitude and then slowly settles into the target

[11] This duration is stipulated in the original theory (Browman & Goldstein, 1989), but becomes a dynamic quantity later.
[12] We will ignore here the idea that plosives have *virtual targets* (Löfqvist & Gracco, 1997).

position. In the simulation we have given, on the other hand, the peak velocity is at the outset. The goal of Articulatory Phonology is not to only achieve a cohesion of phonetics and phonology, but to also be consistent with phonetic nuance. The problem is that in the model all the acceleration happens at the beginning, so a modification to the differential equation that could lead to a slow initial decrease would be to make LA_t be sensitive not only to the distance from LA to LA_0 but also the acceleration of LA, which we will denote by LA_{tt}, as acceleration is the change in velocity with time. The initial acceleration is a large negative number, here, so if we make LA_t be sensitive to the *negative* of LA_{tt}, LA_t would have a large positive boost, slowing down the decrease. The result is the differential equation: $LA_t = -\frac{1}{2}k\,(LA-LA_0) - LA_{tt}$. The factor $\frac{1}{2}$ is there to make this second order system not oscillate when the target has been achieved (Iskarous, 2017), making it *critically damped* (Browman & Goldstein, 1985). This is called a second-order equation since it refers to the change in a change in time, whereas the simplified equation given earlier is a first order equation. A simulation of this second order equation is shown in Figure A3.c., and as seen from the corresponding velocity (bottom row), the very beginning of the action is slow, and then there is a speed up, so the velocity peak is no longer at the very beginning of the movement, which is consistent with the real movement. Improvements on this model are discussed in the main text. If x is any constriction location or degree task variable, the general equation task dynamic equation is usually rearranged and written: $x_{tt} + 2\,x_t + k(x - x_0) = 0$. It is a precise statement about the relation between the phonetic variables x, x_t, and x_{tt} adding up to 0 when weighted by the phonological constants k and x_0, at every moment in time.

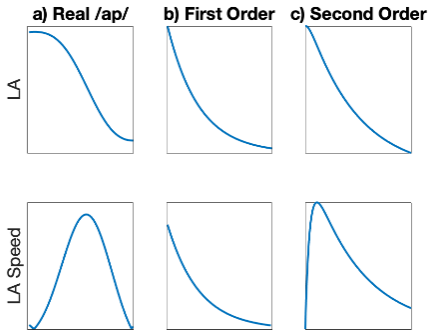

Figure A3. Comparison of LA kinematics of natural utterance (a), first order task equation (b), and second order task equation (c). Top row is task variable and bottom row is magnitude of velocity. The x-axis is time in each case.

Appendix B. Limit Cycle Oscillators

Time can be kept using a system that oscillates in a stable way: If each cycle of oscillation is identical to all others, each cycle's duration, its period, can be taken as a unit of time. A *limit cycle* is such a system. When a stable oscillator is perturbed by some environmental force that tries to change its motion, it will return to its oscillatory state, with the same period and amplitude, without outside help. This is a system with an equilibrium, but not a static equilibrium, instead a dynamic oscillatory one. If we have two oscillators at the same frequency, we can discuss their timing with respect to each other using the notion of *phase*, the

difference in where each of the oscillators is in its cycle of oscillation. Therefore, the limit cycle is a way of keeping time and describing the timing relations of two separate oscillators. The discussion in the following paragraphs will use a mechanical model to explain how a limit cycle works, and it will involve the movements of masses and springs. It may seem that this cannot be something underlying speech, since in the achievement of gestures (other than trills) there is no oscillation. However, neurons and entire brain circuits (Izhikevich, 2007) act as time keeping devices with limit cycle properties in which the oscillation is of the electrical activity of a single neuron and entire neural populations. The mechanical and electrical domains are analogous (Pikovsky et al., 2001), however, the mechanical is easier to visualize. Articulatory Phonology has not been explicit about the specific brain circuits recruited for timing in speech, but believing that such brain circuits could be a part of speech planning does not require a great leap of faith, as other motor systems use such circuits (cf. Poeppel & Assaneo, 2020).

To understand a limit cycle more deeply, it is useful to first understand a simpler, less stable oscillator such as a mass-spring system. Imagine a mass lying on a perfectly smooth table hooked to a spring protruding from a wall. Masses are objects that once in motion stay in motion, and springs are objects that always try to stay at their equilibrium length by having forces that oppose extension or compression. If the spring is not pulled or pushed, the mass will maintain its rest position, with the spring maintaining its equilibrium length. Now if the mass is pulled, a force in the spring is established that tries to bring the mass back to its original position. That force will move the mass back, but since masses in motion stay in motion, the mass keeps moving, squishing the spring and generating a force of extension pushing the mass back. The cycle of extension, motion, and shortening continues endlessly, if the table is perfectly smooth, as can be seen in the in the red curve at the top of Figure B1. This is what is called a linear oscillator, and may already seem like a system that can keep time, as it has equal cycles of oscillation.

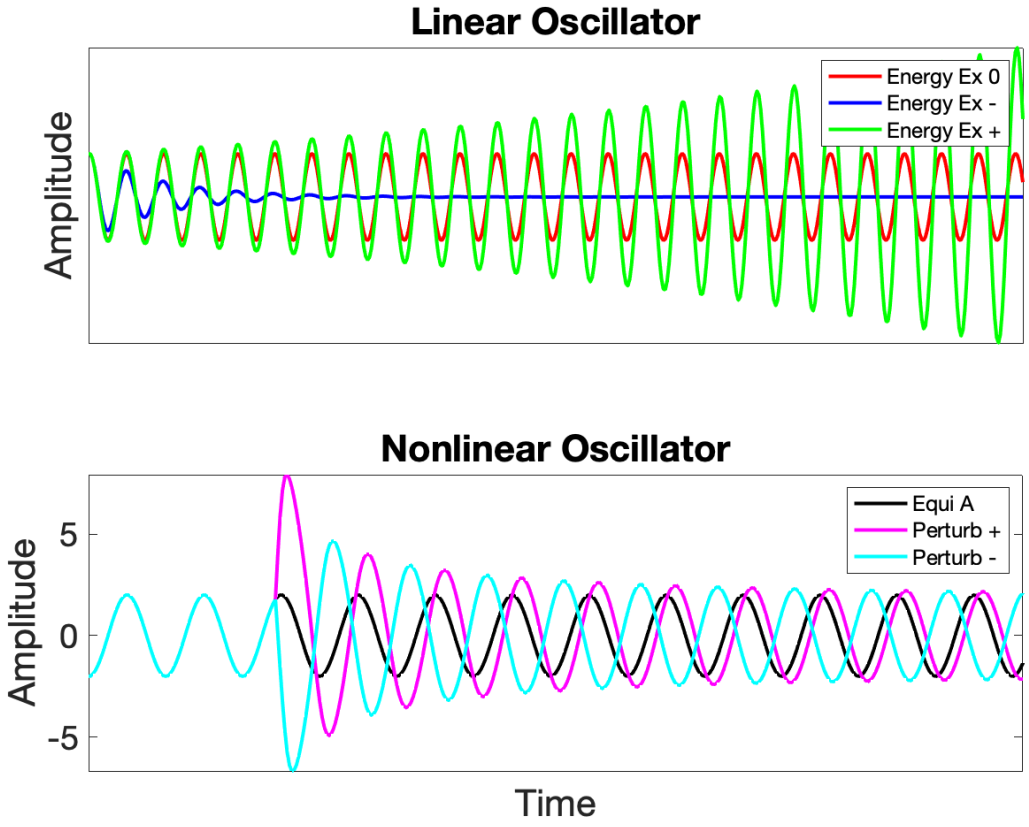

Figure B1. Top panel: Linear oscillators with no exchange of energy with the environment (red), energy-losing (blue), and energy-gaining (green). Bottom panel: Three nonlinear oscillators, all vibrating synchronously, then one's amplitude is suddenly amplified (magenta), another's is suddenly depressed (cyan), and the third is unperturbed (black).

The problem with such a system is that that no surface is perfectly smooth, and friction with the table will make the mass oscillate less and less each cycle as the oscillator comes to a stand-still, as can be seen in the blue curve in Figure B1 (top). Friction makes the particles in the table and mass heat up depositing the energy in the spring onto the vibratory heat of the particles. The linear oscillator is a very poor timer due to friction and energy loss to the environment: it stops ticking.

Imagine, on the other hand, an opposite kind of mass-spring system hooked to a power source that pumps energy into the system from the environment. Instead of the oscillation dying out, more and more energy fed at each cycle will get the oscillator to get further and further away from equilibrium each time, as can be seen in the green curve in Figure B1 (top) till it explodes, making it again less than ideal. So far, we see that imperfection due to either energy sinks (friction) or energy sources (power source) lead to bad timers. What if, however, we can construct a system that oscillates with a particular equilibrium amplitude, and that has a somewhat intelligent interaction with the power source, which is now described. The system oscillates, and if the system loses energy to friction, making its amplitude of oscillation smaller and smaller, the system tunes into the power source, but only under this condition of low amplitude. This power boost eventually restores the original amplitude. But if the amplitude

grows higher than the equilibrium amplitude, the system dumps energy out through friction, leading back to the equilibrium amplitude. In this case, the absorption or shedding of energy is a function of Amplitude of oscillation. Now we come to a definition of a limit cycle: It is a system whose amplitude of oscillation determines its energy interaction with the environment: 1) if its amplitude is at its equilibrium value, there is no energy exchange; 2) if its amplitude is higher than its equilibrium value, the system loses energy; 3) if its amplitude is lower than its equilibrium value, it gains energy.

Figure B1 (bottom) shows three limit cycle oscillators with the same frequency and amplitude, so their curves are coincident initially. The black limit cycle is never perturbed, but at one point in time a perturbation is applied to the cyan and magenta oscillators. The instantaneous perturbation sets the amplitude to a higher value (magenta) or lower value (cyan). In order to get back to the limit cycle, the magenta oscillator severely overshoots the limit cycle amplitude, whereas the magenta oscillation severely undershoots it. And we say that they are stable in their amplitude, with the one that gains energy due to the perturbation shedding it to the environment and returning to its equilibrium amplitude, while the one losing energy due to the perturbation gaining energy and also returning to its equilibrium amplitude. This kind of system is called self-organizing, since its interaction with the environment is not consistent, always losing or gaining energy (and either dying or exploding). Rather, their behavior with respect to the environment is conditional on their current state, making them stable in amplitude. Mathematically this is accomplished through a nonlinear differential equation in which a friction factor depends on the amplitude. Limit cycles may seem like magic, but many inanimate and animate systems in nature behave exactly like that (Pikovsky et al., 2001).

An important observation about limit cycles though (Pikovsky et al., 2001), which can be seen in Figure B1 (bottom) is that even though the amplitude returns to its equilibrium state, the phase of the oscillation is reset by the perturbation. This is called phase-resetting (Pikovsky et al., 2001). Before the perturbation all three limit cycle oscillators are in identical phase, but after the perturbation, the cyan oscillator comes to precede the unperturbed black oscillator, while the magenta oscillator is delayed with respect to the unperturbed black oscillator. This is what we call a *phase difference,* which allows us to define the temporal order of events. Now imagine we have two limit cycle oscillators each oscillating separately. Now introduce a *coupling* between them, making the amplitude of each be sensitive to the amplitude of the other, and start each at a different random phase, so while one is at its peak, maybe the other one has low amplitude. The coupling in amplitude makes the one with the larger amplitude be an energy boosting perturbation on the one with lower amplitude, and vice versa. This bidirectional interaction will lead to both oscillators regaining their equilibrium amplitude, since they are stable in amplitude, but they will change each other's phases, as phase is resettable through perturbation. Under the simplest kind of nonlinear oscillator interaction, the one adopted in Articulatory Phonology, there are two stable phases that such coupled oscillators fall into: In-phase, where the two oscillators are synchronous, and out-of phase, where the two oscillators are negatives of each other. This is the situation for two oscillators. When there are more than two all interacting with each other, other equilibrium phases become possible.

In Articulatory Phonology, each gesture is an oscillator. Syllabic organization is based on the timing of gestures with respect to each other. Language specifies ideal phase differences between gestures as discussed in the main text. Mutual perturbation according to the coupled differential equations of the limit cycles leads to the achievement of stable phase lags or advances, which occurs in a planning stage, before utterances are pronounced. Once a stable phase difference is achieved, the timing difference is used to actually make one gesture start before or after by an amount of time related to the phase lag or advance.

References

Assaneo M. F. & Poeppel D. (2018) The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Sci Adv*. Feb 7;4(2):

Abakarova, D., Iskarous, K., & Noiray, A. (2018). Quantifying lingual coarticulation in German using mutual information: An ultrasound study. *Journal of the Acoustical Society of America*, 144 (2), 897-907.

Amari, S.-I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27, 77–87

Barbosa, P. A. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model for rhythm production. *Proceedings of the Speech Prosody Conference, Aix-en-Provence 2002*, 163-166.

Beckman, M., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody.* (pp. 68-86): Cambridge University Press.

Best, C.T., Goldstein, L. M., Nam, H., & Tyler, M.D. (2016). Articulating what infants attune to in native speech, *Ecological Psychology*, 28, 4, 216-261

Bombien, L. (2011). Segmental and prosodic aspects in the production of consonant clusters. PhD thesis, Ludwig-Maximilians Universität München. https://urn:nbn:de:bvb:19-128407.

Buchaillard, S., Perrier, P., & Payan, Y. (2009). A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning. *The Journal of the Acoustical Society of America,* 126, 2033-2051.

Browman, C. P., & Goldstein, L. (1985). Dynamic modeling of phonetic structure, In V.A. Fromkin (Ed.), *Phonetic Linguistics*, (pp. 35–53), New York, NY: Academic Press,

Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, 45(2–4), 140–155.

Browman, C., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201-251.

Browman, C., & Goldstein, L. (1990). Tiers in Articulatory Phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I. Between the Grammar and Physics of Speech.* (pp. 340-376). Cambridge: CUP.

Browman, C., & Goldstein, L. (1991). Gestural structures: Distinctiveness, phonological processes, and historical change. In *Modularity and the motor theory of speech perception: Proceedings of a conference to honor Alvin M. Liberman* (pp. 313–338). Erlbaum Hillsdale, NJ

Browman, C., & Goldstein, L. (1992). Articulatory Phonology: An overview. *Phonetica, 49*, 155-180.
Browman, C., & Goldstein, L. (1995). Dynamics and Articulatory Phonology. In R. F. Port & T. v. Gelder (Eds.), *Mind as Motion. Explorations in the Dynamics of Cognition.* (pp. 175-194). Cambridge, MA: MIT Press.
Browman, C., & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée, 5*, 25-34.
Bruderer AG, Danielson DK, Kandhadai P, & Werker JF (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, 112(44), 13531–13536.
Brunner, J., Geng, C., Sotiropoulou, S., & Gafos, A. (2014). Timing of German onset and word boundary clusters. *Laboratory Phonology, 5*, 403-454.
Buchaillard, S., Perrier, P. & Payan Y. (2006)A 3D biomechanical vocal tract model to study speech production control: How to take into account the gravity? *Proceedings of the 7th International Seminar on Speech Production*, Ubatuba, Brazil; 2006.
Burroni, F., & Tilsen, S. (in press). The online effect of clash is durational lengthening, not prominence shift: Evidence from Italian. *Journal of Phonetics*.
Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics, 24*, 209-244.
Byrd, D., & Choi, S. (2010). At the juncture of prosody, phonology, and phonetics—The interaction of phrasal and syllable structure in shaping the timing of consonant gestures. In *Papers in Laboratory Phonology X* (pp. 31-60). Berlin: Mouton de Gruyter.
Byrd, D., & Krivokapić, J. (2021). Cracking prosody in Articulatory Phonology. *Annual Review of Linguistics, 7*, 31-53.
Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005). Interacting effects of syllable and phrase position on consonant articulation. *Journal of the Acoustical Society of America, 118*, 3860-3873.
Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics, 26*, 173-199.
Byrd, D., & Saltzman, E. (2003). The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*, 149-180.
Byrd, D., Tobin, S., Bresch, E., & Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *Journal of Phonetics, 37*, 97-110.
Chartier, J., Anumanchipalli, G. K., Johnson, K., & Chang, E. F. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* 98, 1042–1054.
Chen, W., Chang, Y., & Iskarous, K. (2015). Vowel coarticulation: Landmark statistics measure vowel aggression. *Journal of the Acoustical Society of America*, 134, 4167-4189
Cho, T., Kim, D., & Kim, S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics, 64*, 71-89.
Cho, T., Yoon, Y., & Kim, S. (2014). Effects of prosodic boundary and syllable structure on the temporal realization of CV gestures. *Journal of Phonetics, 44*, 96-109.
Choi, D., Dahaene-Lambertz, G., Peña, M., & Werker, J. (2021). Neural indicators of articulator-specific sensorimotor influences on infant speech production. *Proceedings of the National Academy of Sciences*, 118, 20.
Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.
Clements, G. N. (1976). *Vowel harmony in nonlinear generative phonology: an autosegmental model*. Indiana University Linguistics Club.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook, 2*, 225-252.
Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology I* (pp. 283-333). Cambridge: Cambridge University Press.
Clements, G. N. & S. Keyser (1983). *CV Phonology*. The MIT Press
de Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America, 97*, 491-504.
Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *The Journal of the Acoustical Society of America, 89*, 369-382.
Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, 109(3), 545–572
Fant, G. *Acoustic Theory of Speech Production with Calculations Based on X-ray Studies of Russian Articulations*. s'Gravenhag: Mouton, 1960
Fletcher, J. (2010). The prosody of speech: timing and rhythm. In W. J. Hardcastle, J. Laver & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 523-603). Malden, MA: Wiley-Blackwell.
Fowler, C. (1977). Timing Control in Speech Production. PhD Dissertation. Storrs, CT: University of Connecticut.
Fowler, C. (1985). Current perspectives on language and speech production: a critical overview. In R. Daniloff (Ed.), *Speech Science: Recent Advances* (pp. 193-278). San Diego, CA: College Hill Press.
Fowler, C. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14(1), 3–28.
Fowler, C., & Iskarous, K. (2012). Speech production and perception. In I. B. Weiner, A. Healy & R. W. Proctor (Eds.), *Handbook of Psychology. Volume 4: Experimental Psychology* (pp. 236-263): Wiley & Sons.
Fowler, C., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language Production. Volume1: Speech and Talk* (pp. 373-420). London: Academic Press.
Gafos, A. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory, 20*, 269-337.
Gafos, A. (2006). Dynamics in Grammar, in *Laboratory Phonology 8: Varieties of Phonological Competence*, editors M. L. Goldstein, D. H. Whalen, and C. Best (Berlin, New York, NY: Mouton de Gruyter, 51–79.
Gafos, A., Charlow, S., Shaw, J. A., & Hoole, P. (2013). Stochastic time analysis of syllable-referential intervals and simplex onsets. *Journal of Phonetics, 44. 152-166. http://dx.doi.org/10.1016/j.wocn.2013.11.007*.
Gafos, A. , & Benus, S. (2006). Dynamics of phonological cognition. *Cognitive Science* 30, 905–943. doi: 10.1207/s15516709cog0000_80
Gafos, A., & Goldstein, L. (2012). Articulatory representation and organization. In A. Cohn, C. Fougeron & M. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 220-231), Oxford: Oxford University Press.
Gafos, A., & Kirov, C. (2009). A dynamical model of change in phonological representations: The case of lenition. In F. Pellegrino, E. Marsico, I. Chitoran & C. Coupé (Eds.), *Phonological Systems and Complex Adaptive Systems: Phonology and Complexity* (pp. 219-240). Berlin: Mouton der Gruyter.
Galantucci, B., Fowler, C. A., & Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, 71(5), 1138–1149.

Gao, M. (2009). Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. *Chinese Journal of Phonetics, 2*, 43-50.
Georgopoulos, A., Schwartz, A., & Kettner, R. (1986). Neuronal Population Coding of Movement Direction. *Science*, 233, 4771, 1416-1419.
Goldrick, M. (2014). Phonological processing: The retrieval and encoding of word form in speech production. In V. Ferreira, M. Goldrick & M. Miozzo (Eds.), *The Oxford Handbook of Language Production* (pp. 228-244). Oxford: Oxford University Press.
Goldman-Rakic P. S. Circuitry of the primate prefrontal cortex and the regulation of behavior by representational memory. In: Plum F. (Ed.), *Handbook of physiology, the nervous system, higher functions of the brain*. (pp.373-417). Bethesda: American Physiological Society.
Goldsmith, J. (1990). *Autosegmental and metrical phonology*. Basil Blackwell
Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. Arbib (Ed.), *From Action to Language: The Mirror Neuron System* (pp. 215-249). Cambridge: Cambridge University Press.
Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition, 103*, 386-412.
Goldstein, L., & Fowler, C. (2003). Articulatory Phonology: A phonology for public language use. In A. Meyer & N. Schiller (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 159-207). Berlin: Mouton de Gruyter.
Golubitsky, M., Stewart, I., Buono, P L, & Collins, J J (1999). Symmetry in locomoter central pattern generators and animal gaits. *Nature*, 401, 6754, 693-5
Greene, P. H. (1972). Problems of organization of motor systems. *Progress in Theoretical Biology*, 2, 123–145.
Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52, 213–257.
Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. *Progress in Theoretical Biology*, 5, 233–374.
Grossberg, S. (1987). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. *Advances in Psychology*, 43, 313–400.
Gubian, M., Pastätter, M., & Pouplier, M. (2019). Zooming in on spatiotemporal V-to-C coarticulation with functional PCA. *Interspeech, Graz, Austria*, 889-893.
Guitard-Ivent, F., Turco, G., & Fougeron, C. (2021). Domain-initial effects on C-to-V and V-to-V coarticulation in French: A corpus-based study. *Journal of Phonetics, 87*, 101057.
Hermes, A., Mücke, D., & Grice, M. (2013). Gestural coordination of Italian word initial clusters - the case of 'impure s'. *Phonology, 30*, 1-25.
Houde, J. F. & Jordan, M. I. (1998). Sensorimotor Adaptation in Speech Production. *Science*, 279, 5354, 1213-1216.
Houghton G. (1990). The problem of serial order: a neural network model of sequence learning and recall. In Dale R., Mellish C., and Zock M. (Eds) *Current Research in Natural Language Generation.*(pp. 287–319). London: Academic Press.
Iskarous, K. (2005). Patterns of tongue movement. *Journal of Phonetics, 33*, 363-381.
Iskarous, K. (2010). Vowel constrictions are recoverable from formants. *Journal of Phonetics,* 38, 375-387.
Iskarous, K. (2017). The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics, 64*, 8-20.
Iskarous, K. & Kavitskaya, D. (2018). Sound change and the structure of synchronic variability: Phonetic and phonological factors in Slavic palatalization. *Language*. Vol. 94 (1), pp. 4383.

Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., & Whalen, D. H. (2013). The Coarticulation/Invariance Scale: Mutual Information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *Journal of the Acoustical Society of America*, 134, 1271-1284.
Iskarous, K., McDonough, J., & Whalen, D. (2012). A gestural account of the velar fricative in Navajo. *Laboratory Phonology*, 3, 195-210.
Iskarous, K., Nam, H., & Whalen, D. (2010). Perception of articulatory dynamics from acoustic signatures. *Journal of the Acoustical Society of America, 127*, 3717-3728.
Izhikevich, E. (2007). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. The MIT Press.
Jordan, M. I. (1986). Serial Order: A Parallel Distributed Approach. *ICS Report 8604*.
Katsika, A. (2016). The role of prominence in determining the scope of boundary-related lengthening in Greek. *Journal of Phonetics, 55*, 149-181.
Katsika, A., Krivokapić, J., Mooshammer, C., Tiede, M., & Goldstein, L. (2014). The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics, 44*, 62-82.
Karlin, R. (2018). Towards an articulatory model of tone: a cross-linguistic investigation. PhD dissertation, Cornell University. https://doi.org/10.7298/0a3x-9m84
Kelso, Scott (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Bradford Books.
Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14*, 29-59.
Krivokapić, J. (2020). Prosody in Articulatory Phonology. In S. Shattuck-Hufnagel & J. Barnes (Eds.), *Prosodic Theory and Practice*. Cambridge, MA: MIT Press.
Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Phil. Trans. R. Soc. B*.
Krivokapić, J., Styler, W., & Parrell, B. (2020). Pause postures: The relationship between articulation and cognitive processes during pauses. *Journal of Phonetics, 79*, 100953.
Kröger, B. J., Schröder, G., & Opgen-Rhein, C. (1995). A gesture-based dynamic model describing articulatory movement data. *The Journal of the Acoustical Society of America*, 98(4), 1878–1889
Latash, M. (2021). *Bernstein 's Construction of Movements: The Original Text and Commentaries*. Taylor and Francis.
Lammert, A., Goldstein, L., Narayanan, S. and Iskarous, K. (2013). Statistical methods for estimation of direct and differential kinematics of the vocal tract. *Speech Communication*, 55, 147-161.
Lashley, K. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–135). New York: Wiley.
Lee, S., Byrd, D., & Krivokapic, J. (2006). Functional data analysis of prosodic effects on articulatory timing *Journal of the Acoustical Society of America, 119*, 1666 - 1671.
Lee, D. N. (2011*). How movement is guided*. (unpublished manuscript. Retrieved from: http://www.pmarc.ed.ac.uk/ideas/pdf/HowMovtGuided100311.pdf).
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1–36.
Liberman A. M. & Whalen D. H. (2000) On the relation of speech to language. *Trends Cogn Sci*. 4(5):187-196. doi: 10.1016/s1364-6613(00)01471-6.
Liu, Z., Xu, Y., & Hsieh, F.-f. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics, 90*, 101116.

Lloyd, J., Stavness, I. & Fels, S. (2012) ArtiSynth: A Fast Interactive Biomechanical Modeling Toolkit Combining Multibody and Finite Element Simulation. In Y. Payan (Ed.) *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, (pp 355-394), Springer.

Löfqvist, A. , (1990). Speech as audible gestures. In W.J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modeling* (pp. 289–332). Dordrecht: Kluwer Academic Publishers

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

Marin, S., & Pouplier, M. (2014). Articulatory synergies in the temporal organization of liquid clusters in Romanian. *Journal of Phonetics, 42*, 24-36.

Mattingly, I. (1990). The global character of phonetic gestures. *Journal of Phonetics*, 18, 445-452.

McGowan, R., & Saltzman, E. (1995). Incorporating aerodynamic and laryngeal components into Task Dynamics. *Journal of Phonetics, 23*, 255-269.

Meyer, A. (1992). Investigation of phonological encoding through speech error analyses: Achievements, limitations, and alternatives. *Cognition, 42*, 181-211.

Mücke, D., Hermes, A., & Tilsen, S. (2020). Incongruencies between phonological theory and phonetic measurement. *Phonology, 37*, 133-170.

Nam, H. (2007). Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. In J. Cole & J. I. Hualde (Eds.), *Papers in Laboratory Phonology 9* (pp. 483-506). Berlin: Mouton de Gruyter.

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

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In F. Pellegrino, E. Marisco, I. Chitoran & C. Coupé (Eds.), *Approaches to Phonological complexity* (pp. 299-328). Berlin: Mouton de Gruyter.

Nam, H., & Saltzman, E. L. (2003). A competitive, coupled oscillator model of syllable structure. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona

O 'Dell, M., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (Eds.), *Proceedings of th XIVth International Congress of Phonetic Sciences* (Vol. 2, pp. 1075-1078). New York: American Institute of Physics.

O 'Dell, M. L., & Nieminen, T. (2009). Coupled oscillator model for speech timing: overview and examples. *Nordic Prosody: Proceedings of the 10th Conference, Helsinki*, 179-190.

O 'Keefe, J. & L. Nadel (1978). *The Hippocampus is a Cognitive Map*. Oxford: Clarendon Press.

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

Parrell, B. (2021). A potential role for reinforcement learning in speech production. *Journal of Cognitive Neuroscience*, 33, 8, 1450-1486.

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLOS Computational Biology, 15*, e1007321.

Pastätter, M. (2017). The effect of coarticulatory resistance and aerodynamic requirements of consonants on syllable organization. PhD dissertation, LMU Munich Germany. http://nbn-resolving.de/urn:nbn:de:bvb:19-225843

Pastätter, M., & Pouplier, M. (2017). Articulatory mechanisms underlying onset-vowel organization. *Journal of Phonetics, 65*, 1-14.

Pearl, J. (2019) The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM,* 62(3), 54–60. https://doi.org/10.1145/3241036

Perrier, P., Abry, C., & Keller, E. (1988). Vers une modelisation des mouvements de la langue. *Bulletin de la Communication Parlee*, 2, 45–63.
Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). *Synchronization. A Universal Concept in the Nonlinear Sciences*. Cambridge: Cambridge University Press.
Poeppel, D. & Asseano, M. F. (2020). Speech Rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21, 322-334.
Prince, A. & Smolensky, P. (2003). Optimality theory in phonology. *International Encyclopedia of Linguistics*, Vol. 3, 2nd edn., William Frawley (Ed) (pp 212–22). Oxford: Oxford University Press.
Pouplier, M. (2007). Articulatory perspectives on errors. *MIT Working Papers in Linguistics, 53*, 115-132.
Pouplier, M. (2008). The role of a coda consonant as error trigger in repetition tasks. *Journal of Phonetics, 36*, 114-140.
Pouplier, M. (2011). The atoms of phonological representations. In M. van Oostendorp, C. J. Ewen, E. Hume & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 107-129). Malden, MA: Wiley-Blackwell.
Pouplier, M. (2020). Articulatory Phonology. In M. Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press.
Pouplier, M., Hoole, P., & Scobbie, J. (2011). Investigating the asymmetry of English sibilant assimilation: Acoustic and EPG data. *Journal of Laboratory Phonology, 2*, 1-33.
Pouplier, M., & Beňuš, Š. (2011). On the phonetic status of syllabic consonants: Evidence from Slovak. *Journal of Laboratory Phonology*, 2(2), 243-273.
Pouplier, M., & Hoole, P. (2016). Articulatory and acoustic characteristics of German fricative clusters. *Phonetica*, 73, 52-78.
Ramanarayanan, V., Parrell, B., Goldstein, L., Nagarajan, S., & Houde, J. (2016). *Interspeech 2016 Proceeding,* 3564-3568.
Roessig, S., & Mücke, D. (2019). Modeling Dimensions of Prosodic Prominence. *Frontiers in Communication, 4*.
Roon, K. D., & Gafos, A. (2015). Perceptuo-motor effects of response- distractor compatibility in speech: Beyond phonemic identity. *Psychonomic Bulletin & Review*, 22(1), 242–250.
Roon, K., & Gafos, A. (2016). Perceiving while producing: modeling the dynamics of phonological planning. *Journal of Memory and Language, 89*, 222-243.
Saltzman, E. (1979). Levels of Sensorimotor Representation. *Journal of Mathematical Psychology*, 20, 2, 91-163.
Saltzman, E. (1995). Dynamics and coordinate systems in skilled sensorimotor activity. In T. van Gelder & R. Port (Eds.), *Mind as Motion: Dynamics, Behavior, and Cognition* (pp. 149-173). Cambridge, MA: MIT Press.
Saltzman, E., & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science, 19*, 499-526.
Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333-382.
Saltzman, E., Nam, H., Krivokapic, J., & Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In P. A. Barbosa & S. Madureira (Eds.), *Proceedings of the Speech Prosody 2008 Conference, Campinas, Brazil* (pp. 175-184).
Schölkopf, B., Locatello, F., Bauer, S., Ke Nan, R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021) Towards Causal Representation Learning. *Special Issue of Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks* 109(5) 612-634. arXiv:2102.11107.
Schöner, G., & Spencer, J. P.DFT Research Group. (2016). *Dynamic Thinking: A Primer on*

Dynamic Field Theory. Oxford: Oxford University Press.
Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett* (pp. 295-342). Hillsdale, NJ: Lawrence Erlbaum.
Shaw, J. A., & Chen, W.-r. (2019). Spatially Conditioned Speech Timing: Evidence and Implications. *Frontiers in Psychology, 10*.
Shaw, J. A., & Gafos, A. I. (2015). Stochastic Time Models of Syllable Structure. *PLoS One, 10*, e0124714.
Shaw J.A., A.I. Gafos, P. Hoole, & C. Zeroual. (2011) Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. *Phonology*, 28:3, 455-290.
Šimko, J., & Cummins, F. (2010). Embodied Task Dynamics. *Psychological Review, 117*(4), 1229-1246. doi:10.1037/a0020490
Sorensen, T., & Gafos, A. (2016). Autonomous nonlinear gestural dynamics. *Ecological Psychology, 28*, 188-215.
Sotiropoulou, S., Gibson, M., & Gafos, A. (2020). Global organization in Spanish onsets. *Journal of Phonetics, 82*, 100995.
Son, M., Kochetov, A., & Pouplier, M. (2007). The role of gestural overlap in perceptual place assimilation in Korean. In J. Cole & J. I. Hualde (Eds.), *Papers in Laboratory Phonology IX* (pp. 507-534). Berlin: Mouton de Gruyter.
Sorenson, T., Toutios, A., Goldstein, L., & Narayanan, S. (2019). Task-dependence of articulator synergies. *Journal of the Acoustical Society of America*, 145, 3, 1504-1520
Steriade , D. (1990). Gestures and autosegments: Comments on Browman and Goldstein 's paper. In Kingston, J. & Beckman, M. E. (eds.), *Papers in Laboratory Phonology* (1), 382–397
Stetson, R. H. (1951). *Motor phonetics: A study of speech movements in action*. Amsterdam: North-Holland.
Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics, 55*, 53-77.
Tilsen, S. (2019). Motoric Mechanisms for the Emergence of Non-local Phonological Patterns. *Frontiers in Psychology, 10*.
Turk, A., & Shattuck-Hufnagel, S. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*, 193-247.
Turk, A. & Shattuck-Hufnagel, S. (2021). Timing Evidence for Symbolic Phonological Representations and Phonology-Extrinsic Timing in Speech Production. *Frontiers in Psychology*, 10, Article 2952
Turk, A., & Shattuck-Hufnagel, S. (2020). *Speech Timing. Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. Oxford: Oxford University Press.
Turvey, M. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw, & J. Bransford (Eds*.), Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 211–265). Hillsdale, NJ: Erlbaum
Turvey, M. T. (1990). Coordination. *American Psychologist, 45*, 938-953.
Tyrone, M. E., Nam, H., Saltzman, E., Mathur, G., & Goldstein, L. (2010). Prosody and movement in American Sign Language: A task-dynamics approach. In *Speech Prosody 2010* 100957:1-4, http://sprosig.org/sp2010/papers/100957.pdf
Vennemann, T. (1988). *Preference Laws for Syllable Structure and the Explanation of Sound Change*. Berlin: Mouton de Gruyter.
Wilson, H. & Cowan, J. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biophysical Journal*, 12, 1, 1-24.

Yehia, H.C., Kuratate, T., & E. Vatikiotis-Bateson (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics* 30, 555–568.
Yuen, I., Davis, M. H., Brysbaert, M., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences*, 107(2), 592-597.