An informal logic of feedback-based temporal control

Sam Tilsen[1*]

[1]Cornell Phonetics Lab, Department of Linguistics, Cornell University, Ithaca, NY, USA

* Correspondence:
tilsen@cornell.edu

Keywords: articulation, articulatory timing, speech rate, motor control, feedback, dynamical systems, phonology, prosody

Abstract

A conceptual framework and mathematical model of the control of articulatory timing is presented, in which feedback systems play a fundamental role. The model applies both to relatively small timescales, such as within syllables, and to relatively large timescales, such as multi-phrase utterances. A crucial distinction is drawn between internal/predictive feedback and external/sensory feedback. It is argued that speakers modulate attention to feedback to speed up and slow down speech. A number of theoretical implications of the framework are discussed, including consequences for the understanding of syllable structure and prosodic phrase organization.

1 Introduction

Perhaps you have been in a situation in which it was necessary to *shush* someone. For example, imagine you are reading in a library, when a rude person nearby begins talking on their cell phone. You glare at them and say "shhh", transcribed phonetically as [ʃːː]. What determines the duration of this sound? Consider now a different situation: in a coffee shop you are ranting to your friend about the library incident, and your friend tells you to slow down because you are talking too fast. You take a deep breath and proceed more slowly. How do you implement this slowing? The focus of this paper is on how variation in the temporal properties of event durations (your "shhh") and variation in event rate (your rapid coffee shop rant) relate to one another. More specifically, what is the mechanistic connection between control of event timing on short timescales and control of speech rate on longer timescales? It is argued that the answer to this question involves a notion of feedback, and that the same feedback mechanisms are involved on both timescales. In other words, control of event timing involves feedback, and control of rate is reducible to control of timing.

Temporal patterns in speech are challenging to characterize because they exist across a wide range of analysis scales. Figure 1A shows rough approximations of timescales associated with various measurements and theoretical vocabularies. Even over the modest range of 20 ms to 5,000 ms (shown in a logarithmic axis), there is a diversity of ways to associate time intervals with theoretical constructs. Furthermore, there are certain terms—"coordination", "boundaries"—which reappear across scales, and problematically necessitate different interpretations.

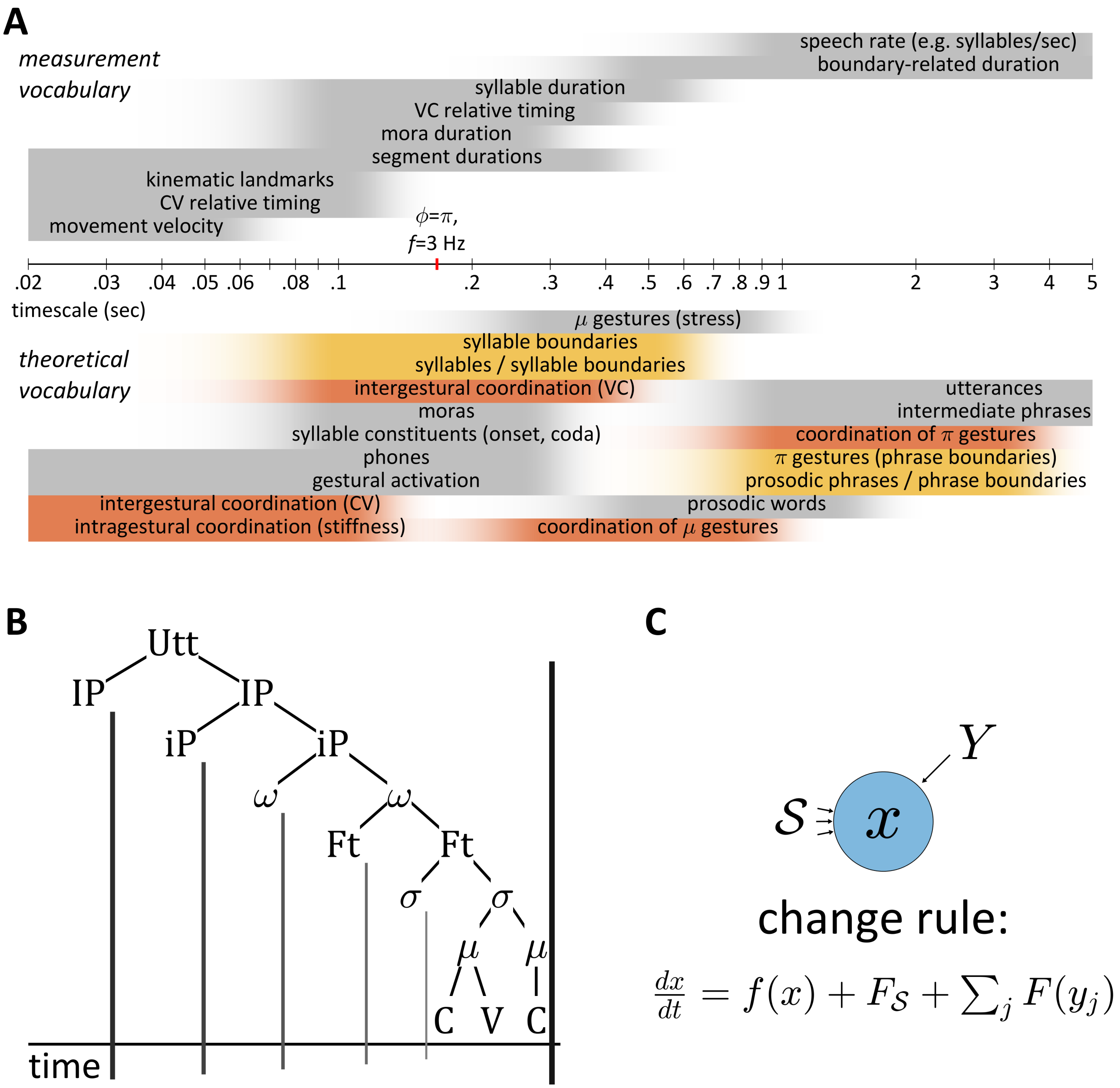

Figure 1. (A) Comparison of timescales associated with various measurements and theoretical constructs used to conceptualize temporal patterns. Time axis is logarithmic. Shaded intervals approximately represent ranges of time in which terminology applied. (B) Hierarchical conception of prosodic structure and implicit projection of units to boundaries in a temporal coordinate. (C) Generic system schema, where change in the state variable *x* is a function of *x* itself and of forces from the surroundings $\mathcal{S}$ and from other systems Y.

It is rarely the case that models of small scale phenomena, such as articulatory timing within syllables, are integrated with models of larger scale phenomena, such as boundary-related slowing. One noteworthy exception is the *π-gesture* model (1), which modulates the rate of a global dynamical clock in the vicinity of phrase boundaries, thereby slowing the timecourse of gestural activation. Another example is the multiscale model of (2), where oscillator-based control of gestural timing is limited to syllable-sized sets of gestures that are competitively selected with a feedback-based mechanism. This early combination of oscillator- and feedback-based control led to the development of Selection-Coordination theory (3,4), an extension of the Articulatory Phonology framework that uses feedback control to account for a variety of cross-linguistic and developmental patterns. A recent proposal in this context is that speech rate is controlled by adjusting the relative contributions of external (sensory) feedback and internal (predictive) feedback (5). One of the aims of this paper is to elaborate on this idea, advancing that generalization that temporal control in speech is largely (but not exclusively) feedback-based.

A broader aim is to argue for a worldview in which speech patterns are understood to result from interactions of dynamical systems. The "informal logic" developed here advocates for new way of thinking about patterns in speech. It is relevant both for the study of speech motor control, specifically in relation to feedback and control of timing, and for theories of phonological representation, sound patterns, and change. The informal logic challenges the prevailing ontologies of many phonological theories by rejecting the notion that speech is cognitively represented as a structure of hierarchically connected objects, as in Figure 1B. It also rejects the notion that such units project "boundaries" onto the temporal dimension of the acoustic signal. Most importantly, the logic holds that speakers never control event durations directly: rather, durational control is accomplished via a class of systems which *indirectly* represent time. They do this by integrating the forces they experience from other systems, or from a surroundings.

The systems-oriented approach can provide a more coherent understanding of temporal phenomena across scales. Its logic is qualified as “informal” because, unlike a formal logic, it does not rely heavily on symbolic forms; rather, the schemas presented below are iconic and indexical, designed to help users rapidly interpret complex patterns of system interactions. At the same time, the schemas can be readily mapped to a explicit mathematical model. All model equations and simulation details are described in Supplementary Material, and all code used to conduct simulations and generate figures has been made available in a repository, here: https://github.com/tilsen/TiR-model.git. Finally, although its implications are fairly general, the scope of this paper is narrowly focused on describing a logic of *temporal* control. Issues related to "spatial" dimensions of feedback or to feedback modalities are set aside for future extensions of the model.

2 Background

In what follows, the objects of our analyses are systems and their relations are interaction forces. Systems are abstract entities which have time-varying internal states. Our analytical task is to formulate change rules to describe how the system states evolve over the course of an utterance, as shown generically in Figure 1C. This setup provides a frame in which to analyze and interpret the causes of empirical patterns in speech. Moreover, to draw generalizations about systems and their interactions we must classify them. To accomplish this in the following sections we define terms below such as *internal*, *external*, *feedback*, and *sensory*. These terms are necessarily relative and therefore potentially ambiguous out of context, thus the reader should pay careful attention to these definitions to avoid confusion.

2.1 Gestural systems and control of gestural activation

Before addressing the role of feedback, we describe the understanding of articulatory control adopted here, which originates from Task Dynamics (6,7). In Task Dynamics (TD), changes in the physical outputs of speech—vocal tract shape and distributions of acoustic energy—are indirectly caused by systems called *articulatory gestures*. Figure 2A schematizes the organization of system interactions in the TD model: gestural systems exert driving forces on vocal tract systems, which in turn exert forces on articulator systems. (As an aside, note that the framework attributes no ontological status to phones or phonemes—these are merely "practical tools" (8) or inventions of scientific cultures (9,10)). Gestural system states are defined in normalized activation coordinates which range from zero to one, and gestures are understood to abruptly become active and subsequently deactivate, as in Figure 2B. When their activation is non-zero, gestures exert forces on vocal tract systems, which can lead to movement, as shown in Figure 2C for timeseries of lip aperture (LA) and pharyngeal constriction (PHAR).

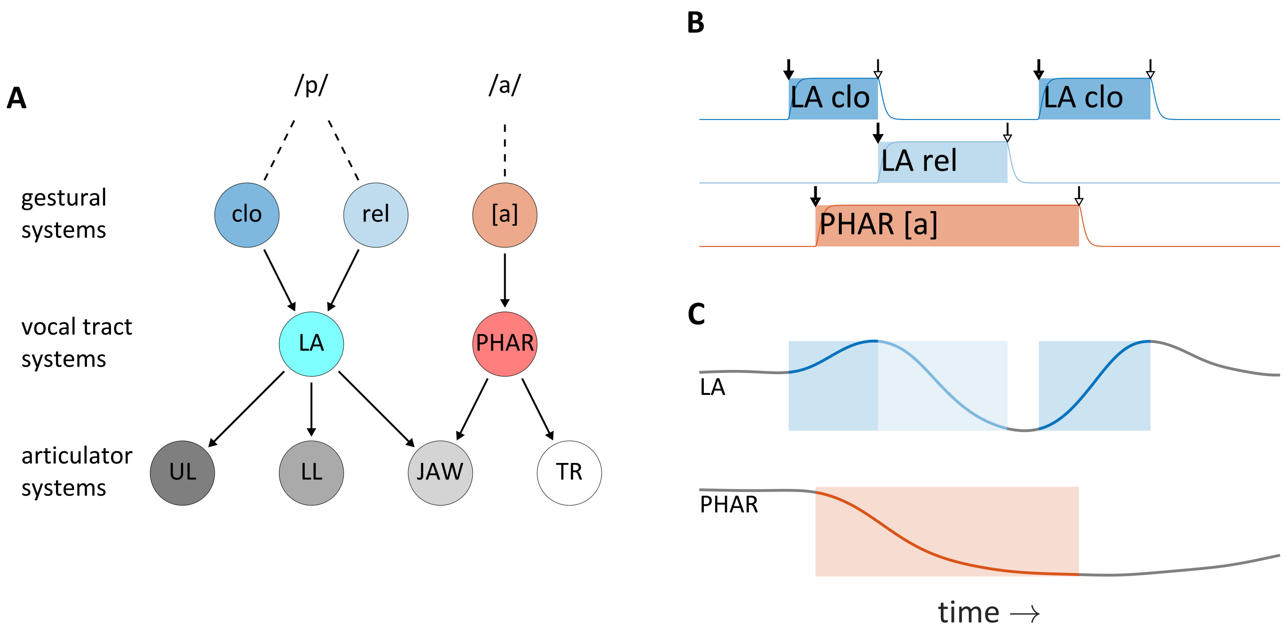

Figure 2. System organization and interactions in the Task Dynamics model. (A) Organization of system interactions. (B) Gestural activation intervals for the CVC syllable *pop*. (C) Vocal tract geometry changes resulting from the actions of gestural systems on vocal tract systems. Lip aperture (LA) and pharyngeal constriction (PHAR) timeseries are shown.

In both a theoretical and technical sense, gestures should be understood as *systems*—entities which have internal states and which experience and exert forces. Accordingly, gestures are not movements, nor are they periods of time in which movements occur. To reinforce this point we often refer to them (redundantly) as *gestural systems*. The distinction is important because it is common to refer to movements of vocal organs as "gestures"—but this can cause confusion. Similarly, the periods in which gestural systems obtain states of high activation (shaded intervals in Figure 2B) are sometimes called "gestures"—these periods are better described as *gestural activation intervals*. The point here is simply that metonymic extensions of "gesture" to refer to physical movements or activation intervals should not be conflated with the systems themselves. Furthermore, the vocal tract and articulator system states of the TD model are nervous system-internal representations of the physical geometry of the vocal tract/effectors. The actual geometry of the vocal tract is not modelled explicitly in TD and can in principle diverge from these internal representations.

The TD framework is particularly valuable because it clarifies the questions that must be addressed in order to understand temporal patterns in speech. There are two questions of paramount importance regarding temporal control: (i) What causes inactive gestural systems to become active? and (ii) What causes active gestural systems to become inactive? These questions are correspond to the arrows marking initiations and terminations of the gestural activation in Figure 2B.

(i) *What causes the gestures to become active?* In answering this question, we temporarily adopt the perspective that the entire set of gestures is a "system". In that case, one possible answer is that there are some *external* systems which exert forces on the gestures. By "external" we mean systems which are "outside" of the set of gestures, and we refer to such systems as *extra-gestural*. Another possibility is that the gestural systems experience forces from each other, in which case the activating forces come from "inside of the system" or are *internal* to the system of gestures, i.e. *inter-gestural*. Note that the first gesture to become active must necessarily be activated by an extra-gestural system, because there is presumably no way for a gestural system to spontaneously "activate itself" or to be activated by inactive gestural systems.

(ii) *What causes the gestures to cease to be active?* The extra-gestural and inter-gestural forces described above are both plausible sources of deactivation. A third possibility, unavailable in the case of activating forces, is that deactivation is caused by actions of individual gestural systems on themselves, i.e. *intra-gesturally*. We elaborate below on how this differs from inter-gestural control.

The Task Dynamics model of speech production developed by Saltzman and Munhall (7) did not resolve which of the various sources of initiating and terminating forces are utilized. Saltzman and Munhall heuristically hand-specified activation intervals to fit empirical data, but they proposed that the model could be extended with the serial network of (11) to dynamically control gestural activation. In this serial network, the hidden layers responsible for sequencing might be interpreted as extra-gestural forces. However, many early descriptions of timing in the TD-based theory of Articulatory Phonology (12,13)—in particular references to "phasing"— imply that initiating forces are inter-gestural and that terminating forces are intra-gestural, in line with the explicit interpretations of phasing in (14). In contrast, later descriptions hypothesize that gestures are activated by a separate system of gestural planning oscillators (15,16), which are extra-gestural.

To summarize, the systems-view of gestural control in the Task Dynamics framework provides two generic options for what causes gestures to become active or cease to be active—extra-gestural systems or other gestures (inter-gestural forces)—along with a third option of intra-gestural control as a form of self-deactivation. There is no theoretical consensus on which of these are actually involved in control of articulatory timing, or in what contexts they may be utilized.

2.2 External feedback vs. internal feedback

The term *feedback* has a variety of different uses. Here *feedback* refers to information which—in either a direct or indirect manner—is produced by some particular system, exists outside of that system, and subsequently plays a role in influencing the state of that same system. Thus feedback is always defined relative to a particular reference system. Feedback in this sense is a very general notion, and does not presuppose that "sensory" organs such as the cochlea or muscle stretch receptors are involved.

For a logic of feedback-based temporal control of speech it is crucial to distinguish between *external feedback* and *internal feedback*, as illustrated in Figure 3. The reference system is the central nervous system (CNS, consisting of cortex, brainstem, and spinal cord). External feedback involves information that (i) is originally generated within the CNS, (ii) is transformed to information outside of the CNS, and (iii) is subsequently transformed back to information within the CNS. For example, activation of the gestural system g_1 causes the production of various forms of information in the environment (movement of articulators, generation of acoustic energy), which is in turn transduced in the peripheral nervous system (depolarization of hair cells in the cochlea and sensory muscle fibers) and subsequently produces information in cortical systems. For current purposes we draw no distinctions between various sensory modalities, which are lumped together as system g'_1 in the Figure 3. The information associated with g'_1 can ultimately influence the state of g_1, and hence meets our definition of feedback. Notice that Figure 3 includes a system labeled $\overline{T}_1$, which uses the external feedback from g'_1 to act on g_1.

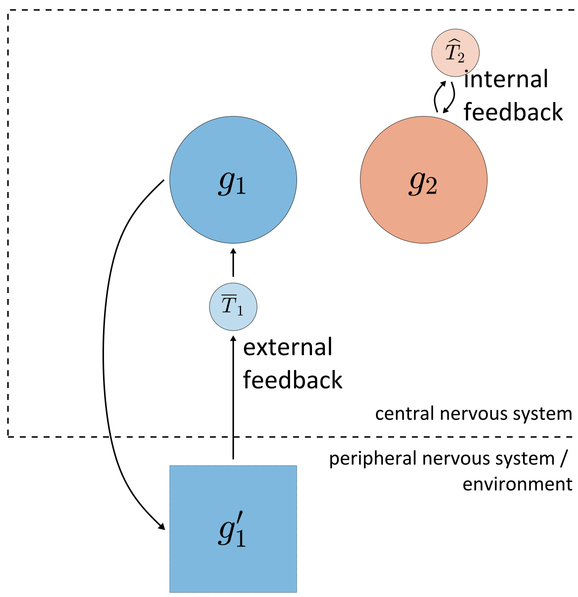

Figure 3. Schematic illustration of distinction between internal and external feedback. The dashed line represents the boundary of the central nervous system. Systems g_1 and g_2 are gestural systems, g'_1 is system which represents information associated with g_1 outside of the central nervous system, and T_1 and T_2 are hypothetical systems which use feedback to act on g_1/g_2.

In contrast to external feedback, internal feedback is information which never exists outside of the CNS. For example, in Figure 3 the gestural system g_2 generates information that system $\widehat{T}_2$ uses to act on g_2. Thus the contrast between external and internal feedback is based on whether the relevant information at some point in time exists "outside of"/"external to" the central nervous system. External feedback may be also described as "sensory" feedback, but with a caveat: one could very well also describe internal feedback as "sensory," in that internal feedback systems experience forces from other systems, and this property can reasonably be considered a form of *sensation*. The point is simply that the word "sensory" is ambiguous regarding what is being sensed, and so the qualifiers *internal* and *external* are preferred, with the CNS being the implied reference system. Internal feedback can also be described as "predictive", but we should be cautious because this term strongly evokes an agentive interpretation of systems.

The distinction between external and internal feedback is only partly orthogonal to distinction between extra-gestural, inter-gestural, and intra-gestural control. The full system of gestures is by definition within the CNS; hence feedback associated with inter-gestural and intra-gestural control is by definition internal feedback. In contrast, extra-gestural control may involve either external feedback (e.g. auditory or proprioceptive information) or internal feedback from CNS-internal systems. This can be confusing because "extra"-gestural control does not entail external feedback—hence the necessity to keep tabs on the system boundaries to which our vocabulary implicitly refers. When describing feedback, the reference system is the CNS. When describing control of gestural activation, the reference system is either the full system of gestures (for extra-gestural control) or individual gestural systems (for inter- vs. intra-gestural control).

The Task Dynamic model incorporates no feedback of any form for gestural systems. Nonetheless, Saltzman and Munhall cited the necessity of eventually incorporating sensory feedback, stating: "without feedback connections that directly or indirectly link the articulators to the intergestural level, a mechanical perturbation to a limb or speech articulatory could not alter the timing structure of a given movement sequence" (8: p. 360). Note that here Saltzman and Munhall expressed a concern with the *temporal* effects of perturbation rather than *spatial* effects—in this paper we are also focused on timing but recognize that a complete picture should incorporate a fully embodied and sensorially differentiated model of the articulatory and acoustic dimensions of feedback.

2.3 Time-representing systems and timing control

To augment our classification of the ways in which gestural systems may be activated or deactivated, we need to think about how time may be "measured", "estimated", or "represented" by the nervous system. Researchers have adopted various ways of talking about different types of systems that serve this function (14,17)—timers, clocks, timekeepers, virtual cycles, etc., with the discussion of (17) being particularly informative. For current purposes, we describe such systems as "time-representers" (TiRs) and develop a multidimensional classification. Despite this name, we emphasize that temporal representations are *always indirect*: the states of the time-representer (TiR) systems are never defined in units of time.

Before classifying TiRs, we make a couple points regarding their interactions with gestures. First, each gestural system is associated with a gating system, labeled "G" in Figure 4A. The gating system states are treated as binary: gates are either open or closed. When a gestural gate is open, the activation state of the associated gestural system transitions rapidly toward its normalized maximum activation of 1. Conversely, when the gate is closed, the gestural system transitions rapidly toward its minimum value. For current purposes, transitions in gestural activation states occur in a single time step, as in (7). Nothing hinges on this simplified implementation and the model can be readily extended to allow for activation ramping or nonlinearities to better fits of empirical tract variable velocity profiles (18).

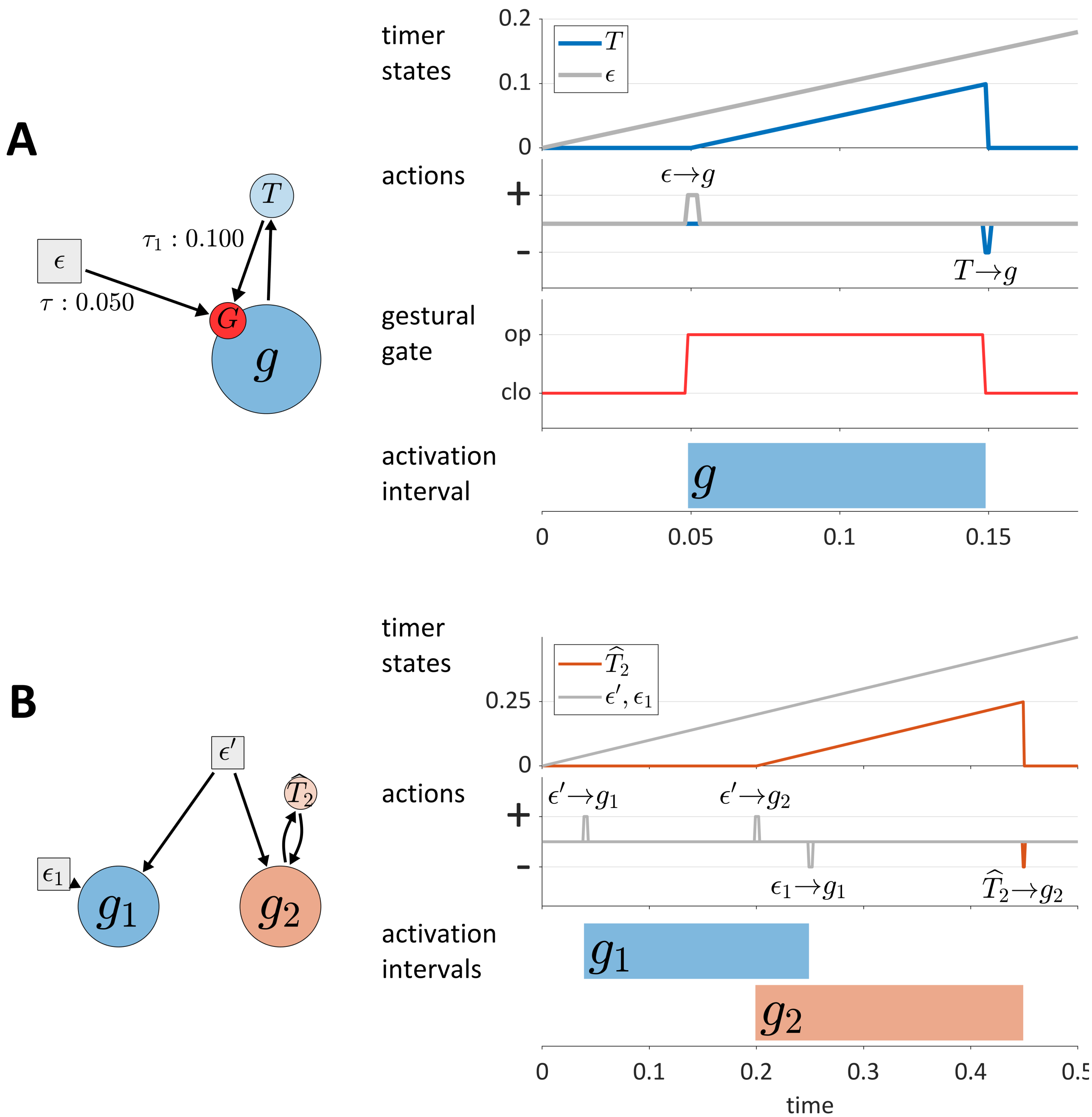

Figure 4. (A) Model of interactions between gestures and TiRs, with depiction of the gestural gating system G that TiRs act upon. Panels on the right show timer states, timer actions on gestures, gestural gating system states, and gestural activation interval. (B) Distinction between autonomous TiRs (ϵ', ϵ_1) and non-autonomous TiRs ($\widehat{T}_2$).

Second, TiRs act on gestural gating systems, not directly on gestures, and thus function to activate/deactivate gestural systems. The actions of TiRs are modeled as brief, pulse-like forces, and always depend on TiR-internal states: each TiR has threshold parameters (τ) which specify the internal states (in units of activation) at which the TiR acts on gating systems. The action threshold parameters are labelled on the arrows of Figure 4A. To reduce visual clutter in model schemas, gating systems are omitted from subsequent figures.

One main dimension of TiR classification involves whether a TiR is autonomous or non-autonomous. An *autonomous* TiR does not depend on either gestural or sensory system input to maintain an indirect representation of time. Figure 4B shows two examples of autonomous TiRs. The first is ϵ', which activates gestures g_1 and g_2. The second is ϵ_1, which deactivates g_1. Note that autonomous TiRs *do* require an external input to begin representing time—they need to be "turned on"/de-gated—but subsequently their state evolution is determined by a growth rate parameter. This parameter may vary in response to changes in a hypothesized "surroundings" or contextual factors.

In contrast to autonomous TiRs, the states of *non-autonomous* TiRs depend on input from a gestural or sensory system. Non-autonomous TiRs integrate the forces that they experience from a given system. An example is $\hat{T}_2$ in Figure 4B, which receives input from g_2 and deactivates g_2 upon reaching a threshold state of activation, here $\tau = 0.25$. Non-autonomous TiRs are associated with integration rate parameters α, which determine how much the forces they experience contribute to changes in their internal states.

The key difference between autonomous TiRs and non-autonomous ones is that the states of the autonomous TiRs evolve independently from the states of gestures or sensory systems. In the example of Figure 4B the states of autonomous TiRs ϵ' and ϵ_1 are assumed to be 0 at the beginning of the simulation and increase linearly in a way that represents the elapsed time. In this example (but not in general), the growth rates of autonomous TiR states were set to $1/\Delta t$, (where Δt is the simulation time step); consequently, their activation states exactly correspond to elapsed time. This is convenient for specifying threshold parameters that determine when TiRs act on other systems. Similarly, the integration rate parameters of non-autonomous TiRs were parameterized to represent the time elapsed from the onset of gestural activation. In general, the correspondence between TiR activation values and elapsed time is neither required nor desirable, and we will see how changes in TiR growth rates/integration rates are useful for modeling various empirical phenomena.

Another dimension of TiR classification involves the sources of input which non-autonomous TiRs make use of to represent time. Non-autonomous TiRs can be described as *external* or *internal*, according to whether they integrate external or internal feedback. This distinction is illustrated in Figure 5A, where the non-autonomous TiR $\hat{T}_1$ can be described as internal because it integrates feedback directly from gesture g_1. In contrast, the non-autonomous TiR $\bar{T}_2$ is external because it integrates feedback from sensory systems which encode the actions of g_2 outside of the CNS.

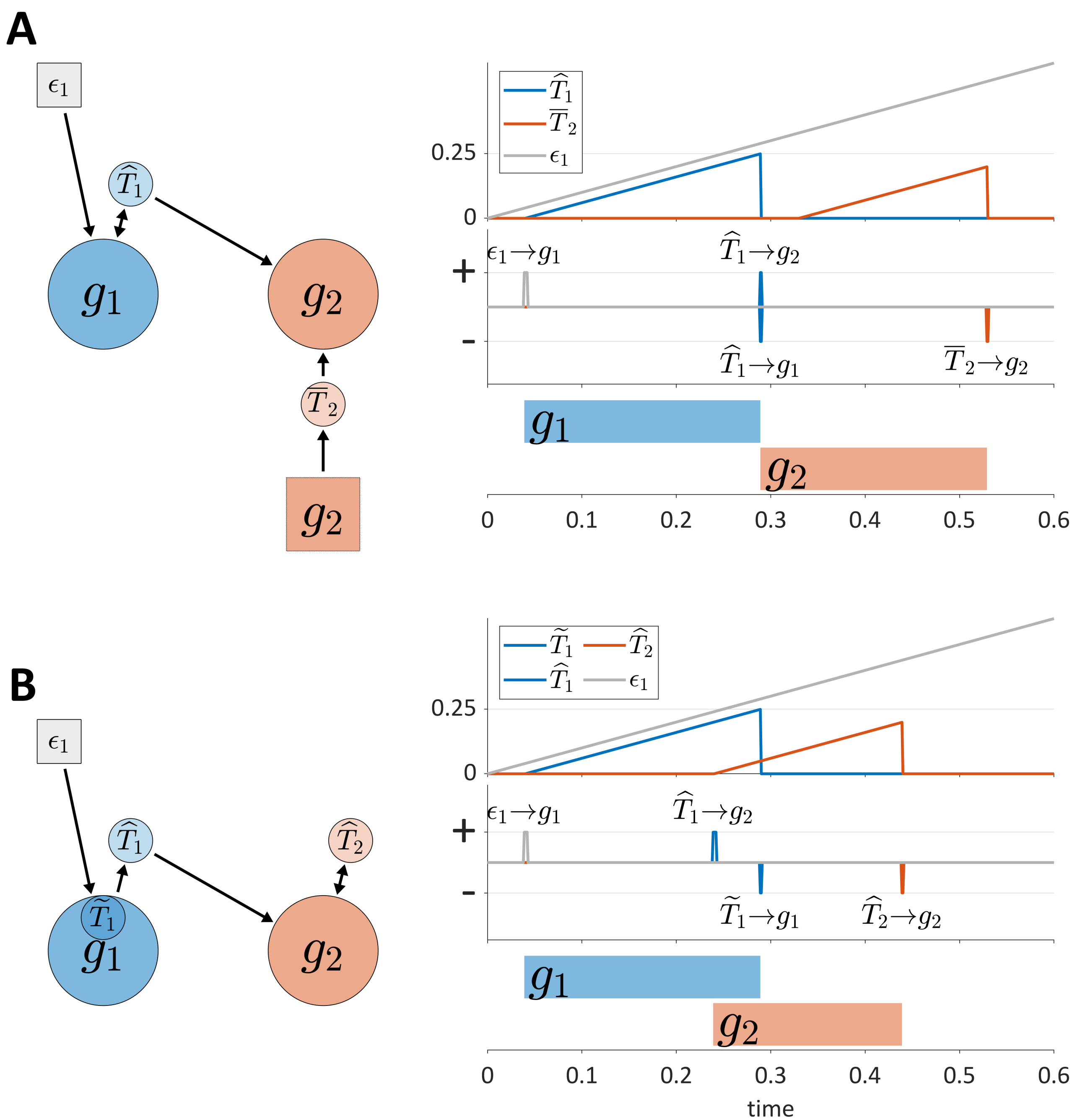

Figure 5. (A) External vs. internal sources of feedback for non-autonomous TiRs. Panels on the right show timer states, timer actions, and gestural activation intervals. (B) Example of inter-gestural vs. isolated/intra-gestural TiRs.

Non-autonomous, internal TiRs are further distinguished according to whether they are inter-gestural or intra-gestural (internal to a gesture). Intra-gestural internal TiRs can only act on the particular gestural system that they are associated with, and can integrate forces only from that gesture. Inter-gestural TiRs can act on and experience forces from any gestural system. For example, in Figure 5B, the deactivation of g_1 is controlled by an intra-gestural TiR $\tilde{T}_1$, but the inter-gestural TiRs $\hat{T}_1$ and $\hat{T}_2$ activate and deactivate g_2, respectively. The distinction is useful if we wish to impose the condition that a TiR is isolated from all systems other than a particular gesture.

The distinction between inter-gestural and intra-gestural TiRs can be viewed in relation to different aspects of the virtual cycles that Tuller and Kelso (14) proposed to govern gestural timing. Tuller and Kelso held that each gesture could be associated with a virtual cycle, which might be described as a "single-shot" oscillation. Different phases of the cycle were hypothesized to correspond to events such as gesture initiation, achievement of maximum velocity, target achievement, and gesture termination. It was suggested in (19) that when a virtual cycle phase of $3\pi/2$ rad (270°) is reached, a gesture is deactivated. In this regard intra-gestural TiRs can implement the functions of virtual cycles: their activation states can be converted to a normalized coordinate that ranges from 0 to 2π, and their growth rates can be adjusted to match the natural frequency of an undamped harmonic oscillator. However, Tuller and Kelso (14) also proposed that intergestural timing might involve specification of the initiation of the virtual cycle of one gesture relative to the virtual cycle of another. Only inter-gestural TiRs can serve this function, because unlike intra-gestural TiRs, they can act on gestural systems that they are not directly associated with. For all of the purposes that follow in this manuscript, intra-gestural TiRs are unnecessary and exclusively use of inter-gestural TiRs.

Autonomous TiRs can differ in whether their state evolution is aperiodic or periodic. Periodic (or technically, quasi-periodic) TiRs are used in the coupled oscillators model (15), where each gesture is associated with an oscillatory system called a *gestural planning oscillator*. The planning oscillators are autonomous TiRs because they do not integrate gestural or sensory system states, as can be seen in Figure 6. They are often assumed to have identical frequencies and to be strongly phase-coupled, such that the instantaneous frequencies of the oscillators are accelerated or decelerated as a function of their phase differences. When a given planning oscillator reaches a particular phase, it "triggers" the activation of the corresponding gestural system. The "triggering" in our framework means that the TiR acts upon a gestural system, in the same way that other TiRs act upon gestural systems. The schema in Figure 6 illustrates a system of three periodic TiRs in which θ_1 and θ_3 are repulsively phase coupled to one another while being attractively phase coupled to θ_2.

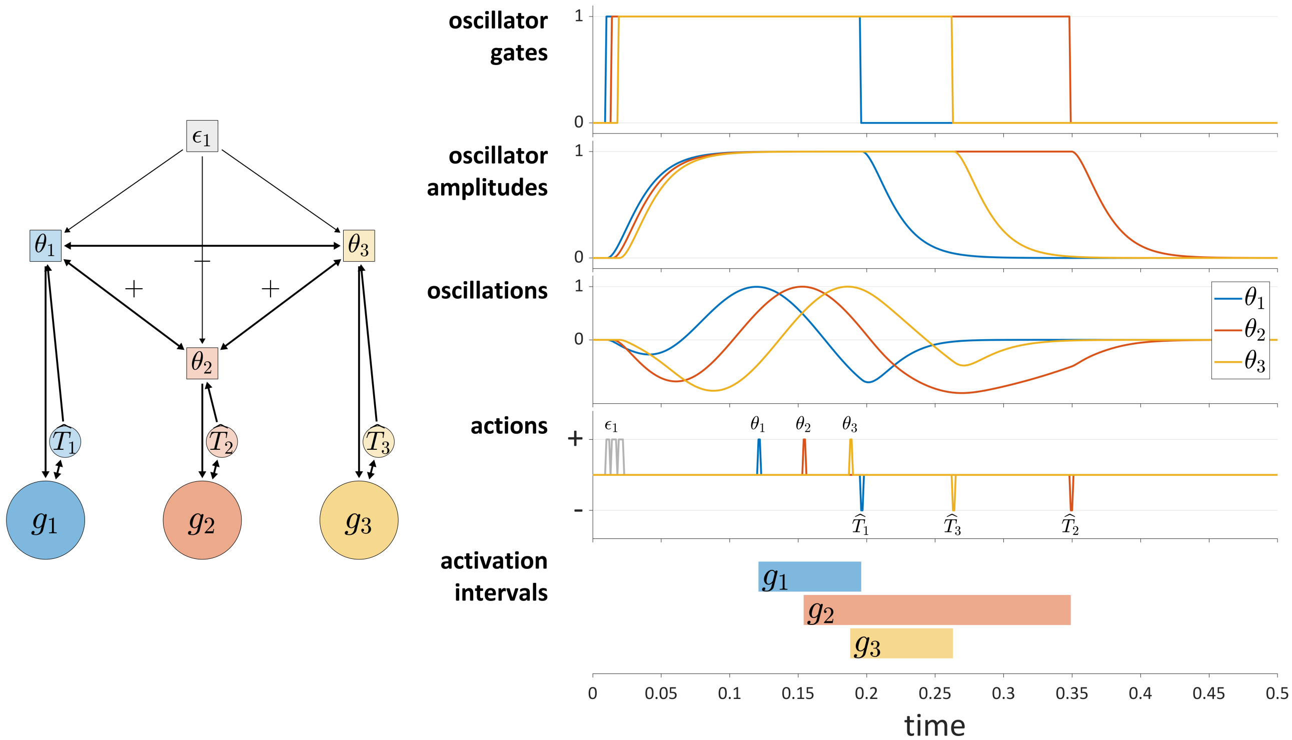

Figure 6. The coupled oscillators model in the TiR framework. Periodic TiRs θ$_1$, θ$_2$, and θ$_3$ are phase coupled as indicated by (+/-) symbols. The oscillator gates, radial amplitudes, and oscillations (amplitude × cosine of phase) are shown. Due to the pattern of phase coupling imposed here, initiation of gestural systems g_1 and g_3 are symmetrically displaced from intiation of g_2.

The phase coupling configuration in Figure 6 generates a pattern of relative phase that—via phase-dependent actions on gestural systems—leads to a symmetric displacement of initiations of gestures g_1 and g_3 relative to initiation of g_2. Statistical tendencies toward symmetric displacement patterns of this sort are commonly observed in two phonological environments: in simple CV syllables, the initiations of constriction formation and release are displaced in opposite directions in time from the initiation of the vocalic gesture (20); in complex onset CCV syllables, the initiations of the first and second constriction are equally displaced in opposite directions from initiation of the vocalic gesture (12,21,22).

The coupled oscillators model has not been used to govern gestural deactivation. Furthermore, a gating mechanism is needed to prevent oscillators from re-triggering gestural systems in subsequent cycles or to prevent them from triggering gestures prematurely. To address this, in the current implementation each oscillator is described by three state variables: a phase angle, a radial amplitude, and the derivative of the radial amplitude. Furthermore, each oscillator is associated with a gating system that controls oscillator amplitude dynamics. These gates are closed by extra-gestural TiRs, as shown in in Figure 6. Moreover, a condition is imposed such that oscillators can only trigger gestural activation when their amplitudes are above a threshold value. The "oscillations" panel of Figure 6 shows a representation of oscillator states that combines phase and amplitude dimensions (the product of the amplitude and the cosine of phase). Further details are provided in the Supplementary Material.

An important hypothesis is that oscillator frequencies are constrained in a way that aperiodic TiR growth rates are not. We refer to this as the *frequency constraint hypothesis*. The rationale is that the oscillator states are believed to represent periodicity in a short-time integration of neuronal population spike-rates; this periodicity is likely to be band-limited due to intrinsic time-constants of the relevant neural circuits and neurophysiology. A reasonable candidate band is theta, which ranges from about 3-8 Hz (23,24), or periods of about 330 to 125 ms. On the basis of these limits, certain empirical predictions regarding temporal patterns can be derived, which we examine in detail below.

Stepping back for a moment, we emphasize that all TiRs can be understood to "represent" time, but this representation is *not* in units of time. The representation results either (i) from the integration of gestural/sensory system forces (non-autonomous TiRs), (ii) from a constant growth rate/frequency (autonomous TiRs) understood to be integration of surroundings forces, or (iii) from a combination of surroundings forces and forces from other TiRs (as in the case of coupled oscillators). Thus the systems we hypothesize represent time indirectly and imperfectly, in units of experienced force.

The utility of TiRs lies partly their ability to indirectly represent time and partly in their ability to act on gestures or other systems. Table 1 below summarizes the types of TiRs discussed above. All TiRs are associated with a parameter vector τ that specifies the activation states at which the TiR acts upon other systems, along with a parameter vector χ whose sign determines whether actions open or close gestural gating systems. Autonomous TiRs are associated with a parameter *ω* which is either a growth rate (aperiodic TiRs) or angular frequency (periodic TiRs). The latter are also associated with a phase-coupling matrix. Non-autonomous TiRs are associated with a vector α of integration factors, which determines how input forces contribute to growth of activation. Additional simulation parameters and details are described in Supplementary Material.

Table 1. Summary of TiRs

symbols	autonomous / non-autonomous	feedback source	sub-classes	periodic/ aperiodic	parameters
ε	autonomous			aperiodic	ω, χ/τ
θ	autonomous			periodic	ω, χ/τ, Φ
$\bar{T}$	non-autonomous	CNS-external	extra-gestural		α, χ/τ
$\hat{T}$	non-autonomous	CNS-internal	inter-gestural		α, χ/τ
$\tilde{T}$	non-autonomous	g-internal	inter-gestural		α, χ/τ

2.4 Deterministic behavior of TiRs and effects of stochastic forces

Under certain conditions, the time δ when a TiR acts on some other system (δ is relative to when TiR activation began to grow) is fully determined by its parameters. In the case of autonomous, aperiodic TiRs, the growth rate ω and action threshold τ determine δ. In two-dimensional ω/τ parameter space, constant δ are straight lines of positive slope, since increases of ω (which shorten δ) can be offset by increases of τ (which lengthen δ). Thus either changes in TiR rate ω or in its action threshold τ, or in some combination of the two, can generate the same change in action timing. This holds for τ and the integration rate α of non-autonomous TiRs as well, as long as the input force to the TiR is constant. For coupled oscillator TiRs, δ depends in complicated ways on the initial phases of the systems, the oscillator frequencies, and the strengths of phase coupling forces (putting aside oscillator amplitude dynamics).

For even a simple system of three gestures, there is a rich set of possible ways in which temporal control can be organized. How can the organization of control be inferred from empirical observations? What we call "noise" may be quite useful in this regard. An essential characteristic of natural speech is that it is unavoidably stochastic, and as a consequence, no two utterances are identical. We interpret stochastic forces here as variation across utterances in the influence of the surroundings on time-representing systems. Moreover, in modeling noise we distinguish between *global noise*—stochastic variation that affects all TiRs equally—and *local noise*—stochastic variation that differentially affects TiRs. This distinction is important because the relative amplitudes of local and global noise can influence timing patterns.

The analysis of stochastic variation below focuses on correlations of successive time intervals between gestural initiations in three-gesture systems. These intervals are referred to as Δ12 and Δ23. We examine correlations (henceforth "Δ-correlations") rather than interval durations, because correlations more directly reflect interactions between systems. Five different local and global noise levels were crossed, from 0 to a maximum level (see Supplementary Material: Simulations for further detail). Figure 7 panels A-F show the structures of each model tested, and corresponding panels A'-F' show how Δ-correlation varies as a function of global and local noise levels. Each line corresponds to a fixed level of global noise, and horizontal points represent different local noise levels.

The "shared trigger" model (A) shows that if both non-initial gestures are activated by feedback from the initial one, Δ-correlation is trivially equal to 1, regardless of noise. The reason for this is simply that the same TiR (here $\hat{1}$) activates g_2 and g_3. Note that this trivial correlation occurs for external feedback control as well (not shown). The coupled oscillators model (B) is unique among the systems examined in that it always produces non-trivial positive correlations. The reason for this has to do with phase coupling. Even when oscillator frequencies are heterogenous due to local noise, phase-coupling forces stabilize the oscillators at a common frequency. As long as phase-coupling forces are strong, local noise has relatively small effects on the phase evolution of oscillators. Global frequency noise always leads to positive correlations because it results in simulation-to-simulation variation in frequency that equally influences Δ12 and Δ23, causing them to covary positively. However, a more complex analysis of correlation structure in the coupled oscillators model in (20) has shown that when coupling strengths are also subject to noise, the model can generate negative correlations.

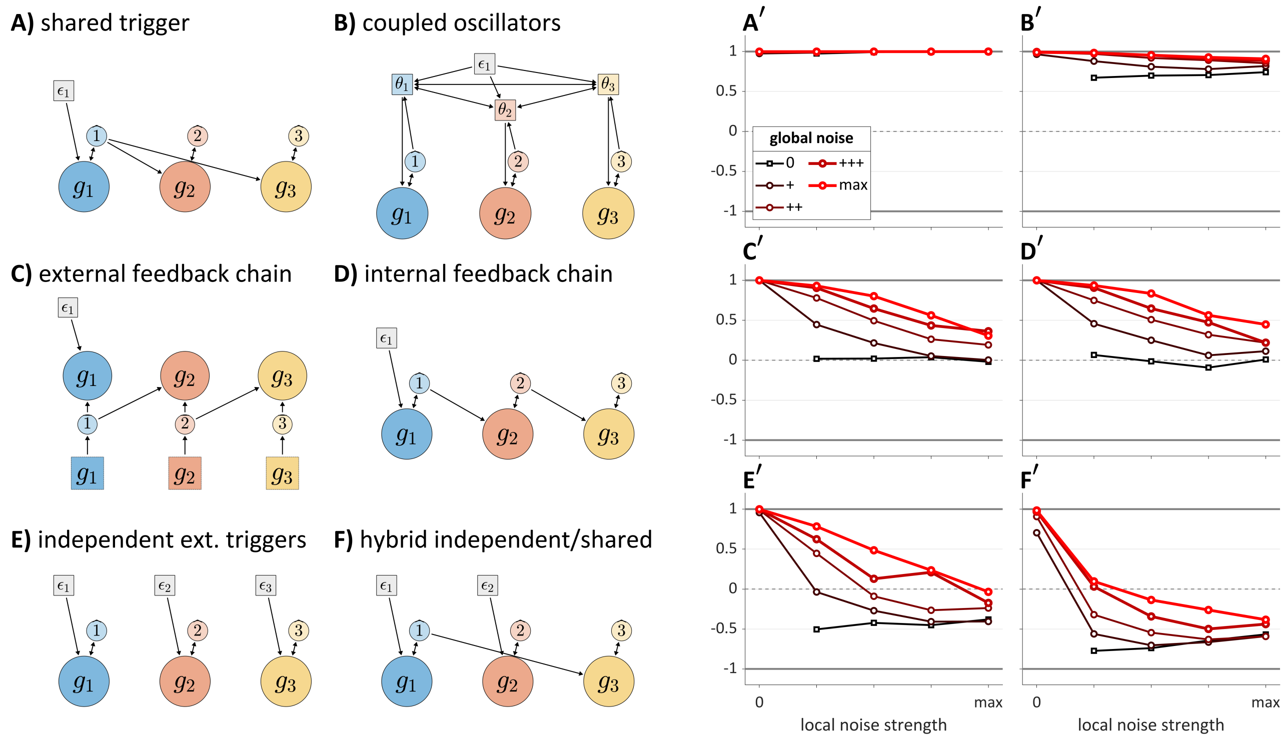

Figure 7. Noise-related correlation patterns for a variety of three-gesture systems. Panels (A-F) show model schemas and corresponding panels (A′-F′) show correlations of intervals between initiation of gestural systems. Local noise levels increase along the horizontal axes, while global noise levels are indicated by the lines in each panel. Cases where both global and local noise are zero are excluded.

The external and internal feedback "chain models" (C and D) exhibit nearly identical, complex patterns of correlation that depend on the relative levels of global and local noise. The patterns are nearly identical because the two models are topologically similar—they are causal chains—differing only in regard to the temporal delay associated with sensory feedback. When there is no local noise, these chain models exhibit Δ-correlations of 1, since the global noise has identical effects on Δ12 and Δ23. Conversely, when there is no global noise, Δ-correlation is 0, since local noise has independent effects on Δ12 and Δ23. In between those extremes, the correlation depends on the relative levels of local and global noise: increasing local relative to global noise leads to decorrelation of the intervals.

Unlike the other models, the independent extra-gestural triggers model (E) and hybrid model (F) can generate substantial negative correlations. In particular, negative correlations arise when g_2 is influenced by local noise. This occurs because whenever the TiR which activates g_2 does so relatively early or late, Δ12 and Δ23 will be influenced in opposite ways. Note that the negative correlations are stronger when the activation of g_1 and g_3 are caused by the same TiR, as is the case for the hybrid model (F). At the same time, global noise induces positive Δ-correlation, counteracting the negative correlating effect of local noise. When we examine speech rate variation below, we will see that the opposing effects of global and local noise are not specific to "noise" per se: any source of variation which has similar effects on all TiRs tends to generate positive interval correlations, while the absence of such variation can lead to zero or negative correlation.

3 A hybrid model of gestural timing and speech rate control

Equipped with a new logic of temporal control, we now develop a hybrid model of gestural timing which is designed to accommodate a wide range of empirical phenomena. The primary requirement of the model is that for each gesture which is hypothesized to drive articulatory movement in an utterance, the model must generate commands to activate and deactivate that gesture.

3.1 Model space and hypotheses

For even a single CVC syllable, the set of all logically possible models is very large. Nonetheless, there are a number of empirical and conceptual arguments that we make to greatly restrict this space. Below we consider various ways in which gestural activation might be controlled for a CVC syllable uttered in isolation. Note that we adopt the modern "split-gesture" analysis in which constriction formation and constriction release are driven by separate gestural systems; this analysis has been discussed and empirically motivated in (20,25,26). With that in mind we use the following gestural labeling conventions: C/c and R/r correspond to constriction formation and release gestures, respectively; upper case labels C/R correspond to pre-vocalic gestures (or, gestures associated with syllable onsets); lower case labels c/r correspond to post-vocalic gestures (or, gestures associated with syllable codas); and gestures/gesture pairs are subscripted according to the order in which they are initiated.

The schemas in Figure 8 (A-C) show "extreme" models which—though logically possible—are conceptually and empirically problematic. (A) shows a "maximally sensory" model, where all gestural activation/deactivation is controlled by external feedback systems. This model is problematic because the time delay between efferent motor signals and afferent feedback is too long to be useful for some relative timing patterns, such as the relative timing of consonantal constriction and release in normal speech. (B) shows a "maximally internal" model, where all gestural activation and deactivation is induced by inter-gestural TiRs (keeping in mind that initiation of activation of the first gesture in an utterance is always external). The maximally internal model is problematic because it has no way of allowing for external/sensory feedback to influence timing.

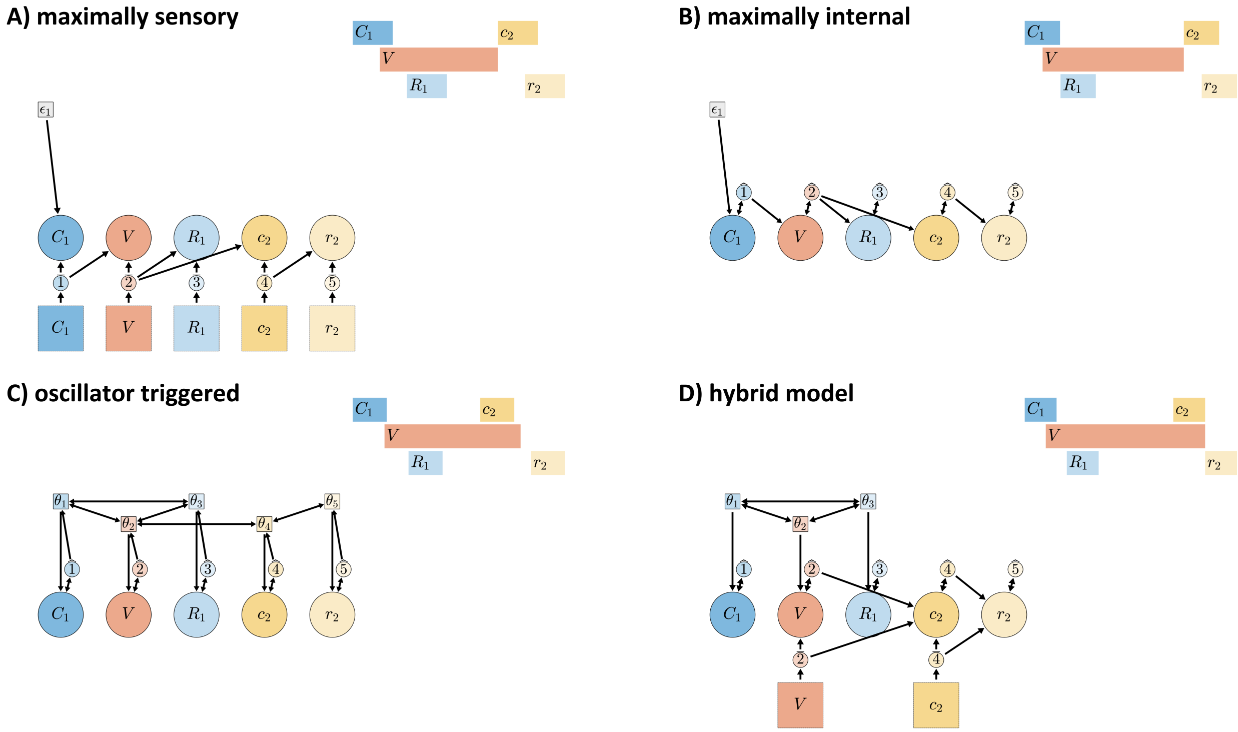

Figure 8. Candidate models of CVC syllables. (A) Maximally sensory model where all activation and deactivation is controlled by external sensory feedback. (B) Maximally internal model where all control is governed by internal feedback. (C) Fully oscillator-triggered model where all gestures are initiated by oscillators. (D) Hybrid model in which pre-vocalic gestural activation is oscillator-governed while post-vocalic activation is governed by either internal or external feedback.

Schema (C) shows an "oscillator triggered" model, where all gestures are activated by coupled oscillators. Under standard assumptions, this model is problematic because it cannot generate some empirically observed combinations of pre-vocalic and post-vocalic consonantal timing, as discussed in (5). The "standard" assumptions are: (i) that all oscillators have (approximately) the same frequency; (ii) that all oscillators trigger gestural initiation at the same phase of their cycle; and (iii) that only in-phase and anti-phase coupling are allowed. With these constraints, the model cannot generate empirically common combinations of pre-vocalic and post-vocalic temporal intervals, where prevocalic CV intervals are generally in the range of 50-100 ms (20) and post-vocalic VC intervals—periods of time from V initiation to post-vocalic C initiation—are in the range of 150-400 ms. Moreover, relaxing any of the three assumptions may be undesirable. Allowing oscillators to have substantially different frequencies can lead to instability and chaotic dynamics, unless coupling forces are made very strong. Allowing oscillators to trigger gestures at arbitrary phases is inconsistent with the neurophysiological interpretation: presumably one particular phase of the cycle represents maximal population spike rate and should be associated with the strongest triggering force. Allowing for arbitrary relative phase coupling targets, such as a relative phase equilibrium of $3\pi/2$, may not be well-motivated from a behavioral or neurophysiological perspective.

Although the relatively extreme/monolithic models of Figure 8 (A-C) are individually problematic, the mechanisms that they employ are practically indispensable for a comprehensive understanding of timing control. External feedback control is necessary to account for common observation that segmental durations are lengthened in the presence of feedback perturbations (27–32). Internal feedback is necessary to allow for control under circumstances in which external feedback is not available, for example during loud cocktail parties, for speakers with complete hearing loss, or during subvocal rehearsal (internal speech) with no articulatory movement. Finally, oscillator-triggered control is currently the only known mechanism which adequately explains symmetric displacement patterns (5,20). Given the utility of these mechanisms it is sensible to adopt a hybrid model which combines them, as in Figure 8D. The hybrid model of (D) represents the following two hypotheses.

Pre-vocalic coordinative control hypothesis. Control of the activation of pre-vocalic consonantal constriction formation (C), release (R), and vocalic initiation (V) is governed by a system of coupled oscillators.

Vocalic/post-vocalic feedback control hypothesis. The deactivation of vowel gestures and the activation/deactivation of post-vocalic constriction (c) and release (r) gestures is governed by either internal or external feedback.

Together these hypotheses are referred to as the *hybrid control model*. The specific predictions of the hypotheses are best considered in light of how interval durations change in response to other sources of variation, which we examine below.

3.2 External influences on parameters

The parameters of TiRs are context-dependent: they vary in ways that are conditioned on factors associated with their surroundings, so-called "external factors". Here we demonstrate two ways in which external factors may influence timing. An innovation of the model is the idea that these factors can have differential influences on external vs. internal TiR parameters.

Figure 9 (A-C) demonstrates the effects of variation in a hypothetical contextual factor of *self-attention*, or "attention to one's own speech". The figure summarizes simulations of the system shown in panel (A), where activation of a post-vocalic constriction gesture c_1 is potentially caused by an internal or external TiR representing feedback from the vocalic gesture V_1. This is the hypothesized organization of post-vocalic control in the hybrid model. An external variable λ is posited to represent self-attention. By hypothesis, the force integration rates of internal and external TiRs are differentially modulated by λ, such that $\alpha = \alpha' / (1 + \beta\lambda)$, where $\beta_{internal} < \beta_{external}$. This reflects the intuition that when one attends to feedback more closely, feedback-accumulation (i.e. force-integration) rates of TiR systems are diminished, so that TiRs take longer to act on gestures. This diminishing effect applies more strongly to internal feedback than external feedback. As a consequence, there is a value of λ such that as λ is increased, initiation of g_2 switches from being governed by the internal TiR to the external one. In the example the transition occurs around λ = 0.425, where a change is visible in the slope relating the control parameter λ and the interval δ (the time between initiation of V_1 and c_1). Gestural activation intervals associated with three values of λ are shown in panel (C).

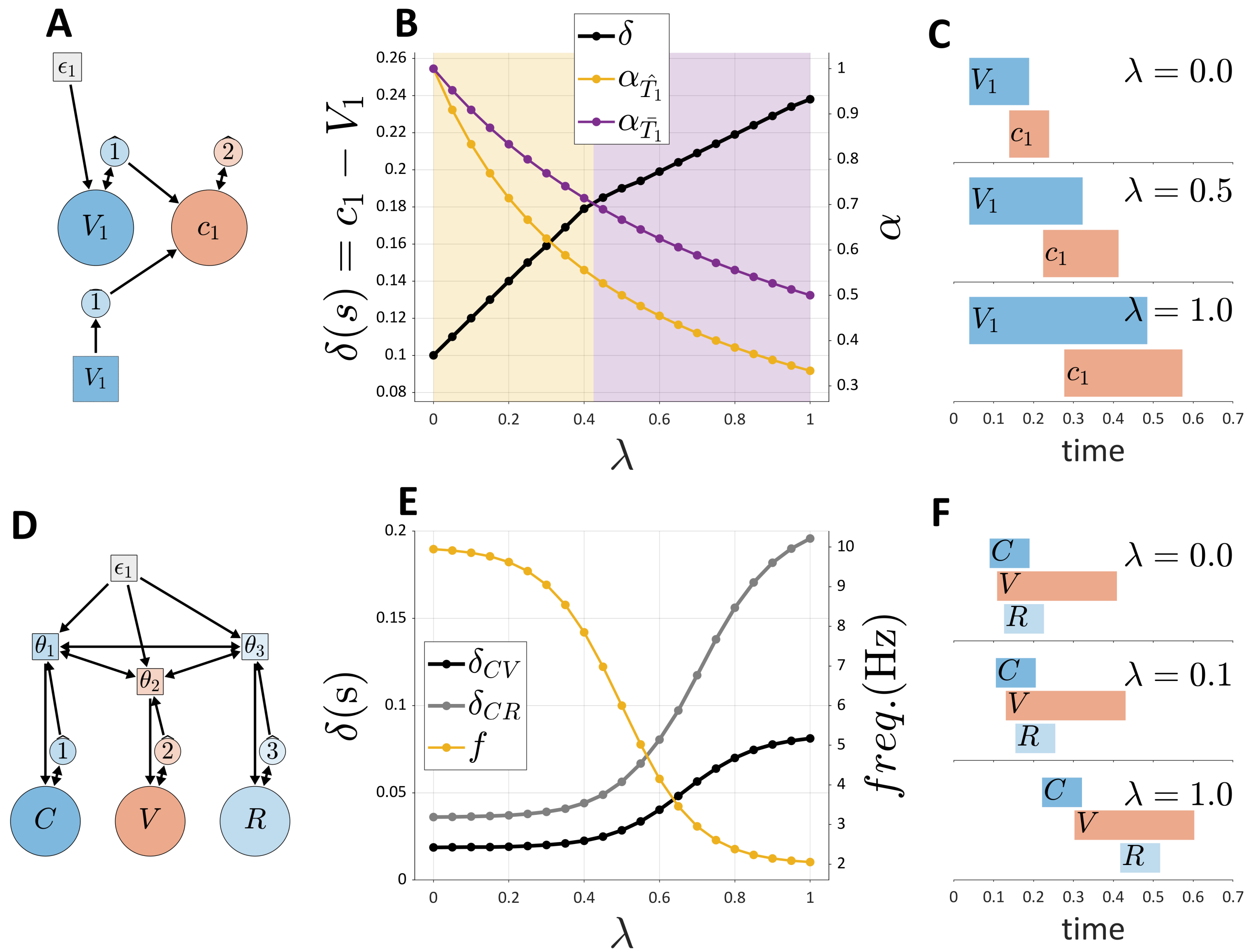

Figure 9. Simulations of external influences on parameters. (A) Schema for post-vocalic control with both internal and external TiRs. (B) Dual axis plot showing how δ (left side) and integration rates α (right side) change with self-attention parameter λ. (C) Gestural activation intervals for several values of λ. (D) Model schema of pre-vocalic coordinative control. (E) Dual axis plot showing effect of rate parameter λ on δ-values (left side) and frequencies (right side). (F) Gestural activation intervals for several values of λ.

Panel (B) shows that when TiR parameters are differentially modulated by an external influence, transitions between internal and external feedback control can occur. In the above example, the external influence was posited to represent "self-attention" and its state was encoded in the variable λ; this variable was then hypothesized to differentially adjust external vs. internal non-autonomous TiR growth rates. An alternative way in which the same effect might be derived is by allowing the external variable λ to differentially adjust TiR action-thresholds. Realistically, external variables of this sort may influence both growth rate and threshold parameters.

Another parameter that can respond to external factors is the frequency of the coupled oscillators which are hypothesized to govern prevocalic gestural initiation. Suppose that the external factor here is a something novel that we call "pace" and that pace influences oscillator frequencies. However, because of the frequency constraint hypothesis, we cannot simply allow the oscillator frequencies to respond linearly to changes in pace. Instead, we impose soft upper and lower frequency bounds by attenuating the effect of the pace parameter λ on frequency f. This is accomplished by making the effective frequency a nonlinear function of λ, as shown in Figure 9E (right side). The consequence of this limitation on f is that intervals which are governed by coordinative control are predicted to exhibit nonlinear responses to variation in the external factor: here we can see that the δCV and δCR plateau at extreme values of λ.

In section 3.4 we combine the above effects of self-attention and pace into a general model of the control of speech rate. But first we introduce another important mechanism, which allows the model to organize the subsystems of larger utterances.

3.3 Parallel domains of competitive selection

Competitive selection (or competitive queuing) is a dynamical mechanism that, given some number of actions, iteratively selects one action while preventing the others from being selected. The concept of competitive selection of actions originates from (33), and many variations of the idea of have been explored subsequently, both within and outside of speech (2,34–39). One of the key ideas behind the mechanism is that a serial order of actions is encoded in an initial activation gradient, such that prior to the performance of an action sequence, the first action in the sequence will have the highest relative activation gradient, the second action will have the next highest activation, and so on. The growth of activation is a "competition" of systems to be selected, and selection is achieved by reaching an activation threshold. Moreover, action selection is mutually exclusive, such that only one action can be selected at a time.

Figure 10 shows how these ideas are understood in the current model. The "actions" which are competitively selected in this example are three CV syllables, and the selection of these actions is governed by systems that we refer to as *μ-systems*. As shown in the model schema, each μ-system de-gates a system of coupled oscillators, which in turn activate gestures. Each of the μ-systems is associated with a μ-gating system that—when open—allows the corresponding μ-system activation to grow. Notice that at time 0 (before the production of the sequence), the pattern of relative activation of μ-systems corresponds to the order in which they are selected. When μ-system gates are open, μ-system activations grow until one of the systems reaches the selection threshold. At this point, all μ-gating systems are closed, which halting growth of μ-system activation. The selected μ-system is eventually suppressed (its activation is reset to 0) by feedback—specifically by the inter-gestural TiR associated with the last gesture of the syllable, in this case the vowel gesture. This causes all μ-systems to be de-gated, allowing their activations to grow until the next most highly active μ-system reaches the selection threshold. This three-step process—(i) de-gating and competition, (ii) selection and gating of competitors, and (iii) feedback-induced suppression of the selected system—iterates until all of the μ-systems have been selected and suppressed. See Supplementary Material: Model details for further information regarding the implementation.

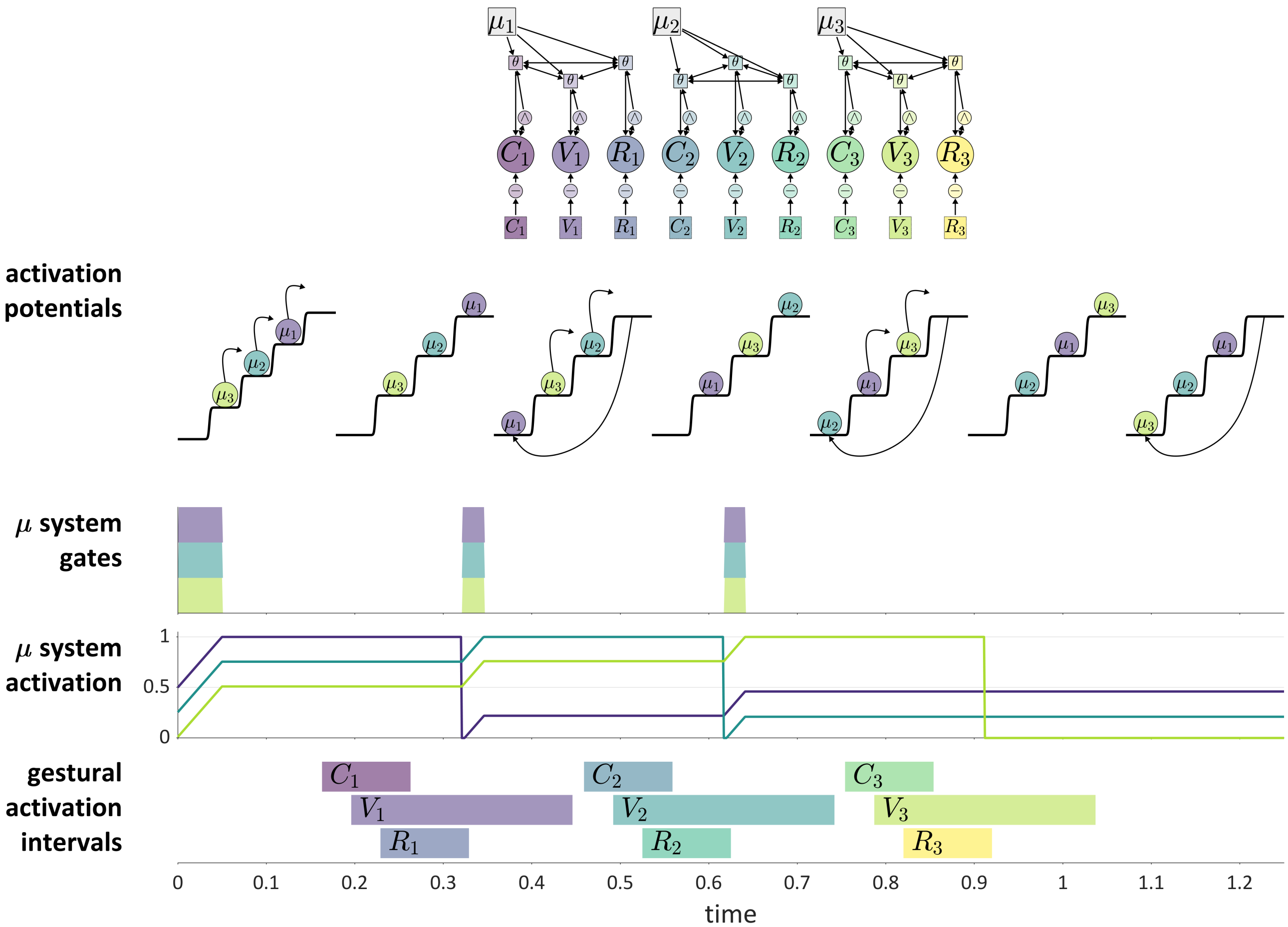

Figure 10. Illustration of competitive selection for a sequence of three CV syllables. Top: model schema. Activation potentials with arrows show transitions between states, and potentials without arrows shown quasi-steady states. μ-gating system states are shown (shaded intervals are open states). Bottom: gestural activation intervals.

A more abstract depiction of a competitive selection trajectory is included in the activation potentials of Figure 10. The potentials without arrows are relatively long epochs of time in which μ-systems exhibit an approximately steady-state pattern of activation. The potentials with arrows correspond to abrupt intervening transitions in which the relative activation of systems is re-organized by the competitive selection/suppression mechanism. Along these lines, the dynamics of competitive selection have been conceptualized in terms of operations on discrete states in (40,41).

There are two important questions to consider regarding the application of a competitive selection mechanism to speech. First, exactly what is responsible for suppressing the currently selected μ-system? In the example above, which involves only CV-sized sets of gestures, it was the internal TiR associated with the last gesture of each set. Yet a more general principle is desirable. Second, what generalizations can we make about the gestural composition of μ-systems? In other words, how is control of gestural selection organized, such that some gestures are selected together (*co-selected*) and coordinatively controlled, while others are competitively selected via feedback mechanisms? This question has been discussed extensively in the context of the Selection-coordination theory of speech production (3–5), where it is hypothesized that the organization of control follows a typical developmental progression. In this progression, the use of external sensory feedback for suppression/de-gating is replaced with the use of internal feedback, a process called *internalization of control*.

The are two important points to make about internalization. First, internalization of control is partly optional, resulting in various patterns of cross-linguistic and inter-speaker variation which are detailed in (3) and which we briefly discuss in section 4.1. Second, internalization is flexible within and across utterances, such that various contextual factors (e.g., self-attention) can influence whether external or internal feedback TiRs are responsible for suppressing selected μ-systems.

Furthermore, a recently developed theory of syntactic organization in speech (40) argues that there are two interacting domains of competitive selection. This is known as the *parallel domains hypothesis*. One of these domains involves "gestural-motoric" organization of the sort illustrated above, where gestures are organized into competitively selected sets (μ-systems). The other involves "conceptual-syntactic" organization in which concept systems are organized into competitively selected sets. The hypotheses advanced in (40) hold that sets of co-selected conceptual systems correspond loosely to the prosodic unit called the *phonological word* (a.k.a. pwrd, or ω), which has the property that there is a single accentual gesture associated with set of co-selected conceptual systems. Moreover, under normal circumstances speakers do not interrupt (for example by pausing) the gestural competitive selection processes which are induced by selection a phonological word.

These parallel domains of conceptual-syntactic and gestural-motoric competitive selection are illustrated Figure 11 for an utterance which would typically be analyzed as four prosodic words, such as [*a dog*] [*and a cat*] [*chased*] [*the monkey*]. Note that to conserve visual space release gestures have been excluded. The top panel shows the sequence of epochs in competitive selection of concept systems $\mathcal{C}$. Each of these could in general be composed of a number of co-selected subsystems (not shown). For each epoch of concept system selection, there is a corresponding series of one or more epochs of competitive selection of gestural systems. The model accomplishes this by allowing the concept systems to de-gate the corresponding sets of μ-systems. Within each of these sets of μ-systems, the appropriate initial activation gradient is imposed. Further detail on the implementation is provided in the Supplementary Material.

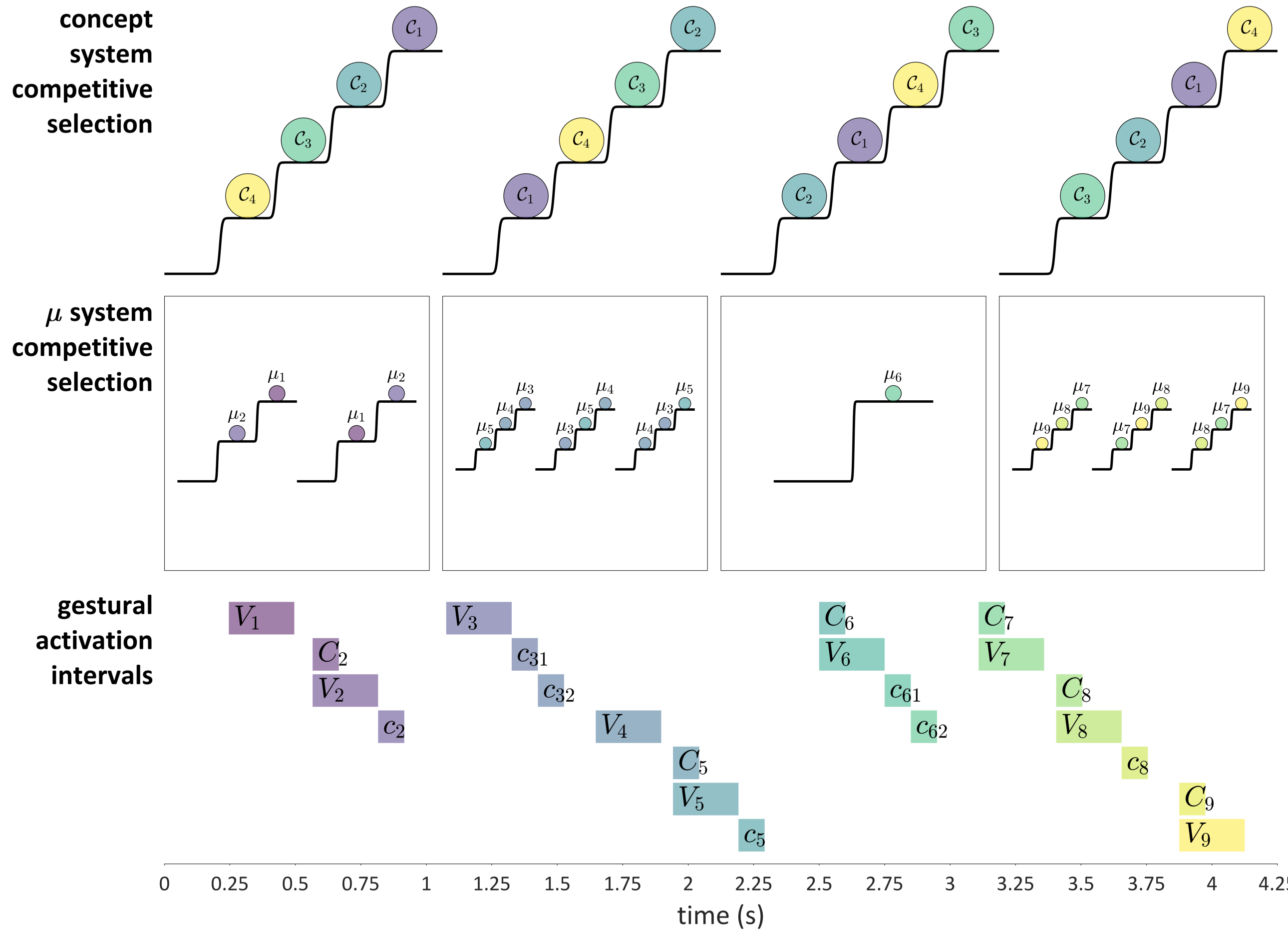

Figure 11. Illustration of parallel domains of competitive selection for an utterance with the structure. Top: concept systems $\mathcal{C}$ are competitive selected. Middle: selection a concept system de-gates corresponding μ-systems which themselves are competitively selected. Bottom: gestural activation intervals generated by the model.

Although there is no *a priori* constraint on the number of domains of competitive selection that might be modelled, the parallel domains hypothesis that we adopt makes the strong claim that only two levels are needed—one for conceptual-syntactic organization and one for gestural-motoric organization. We examine some of the important consequences of these ideas in section 4.2, regarding phrasal organization. One aspect of prosodic organization which we do not elaborate on specifically in this paper involves the metrical (stress-related) organization of gestures, but see (42) for the idea that the property of "stress" relates to which sets of co-selected gestures (μ-systems) may include accentual gestures, which in turn are responsible for transient increases in self-attention.

3.4 A model of speech rate control with selectional effects

When given verbal instructions to "talk fast" or "talk slow", speakers are able to produce speech that listeners can readily judge to be relatively fast or slow. To quantify this sort of variation, speech rate is often measured as a count of events per unit time, e.g., syllables per second or phones per second. There are several important points to consider about these sorts of quantities. First, in order to be practically useful, an event rate must be measured over a period of time in which multiple events occur. As the size of the counting window decreases, eventually only one full event is included.

Second, there is no consensus on which events are the appropriate ones to count—phones, syllables, words, or something else? In the current framework, many commonly used units do not even have an ontological status. Third, even if we ignore the above problems, the resulting rate measure cannot be assumed to be a very good reflection of what speakers are controlling at any particular instant. There is no evidence to my knowledge that speakers directly control rate quantities such as syllables/second or phones/second. If we infer that speakers do not in fact control speech rate as an event rate *per se*, then what are speakers controlling in order to speak fast or slow?

The *attentional modulation hypothesis* (5) holds that speakers control rate by modulating their attention to feedback of their own speech (*self-attention*), and specifically do so in a way that, as self-attention increases, prioritizes external/sensory feedback over internal feedback. Furthermore, this hypothesis holds that along with modulating self-attention, speakers may adjust pacing, that is, the frequencies of gestural planning oscillators. The separate effects of varying these external factors were already demonstrated in Section 3.2.

In addition, a mechanism is need to account for the phenomenon of boundary-related lengthening. Many empirical studies have shown that speech slows down as speakers approach the ends of phrases, with greater slowing and increased likelihood of pausing statistically associated with "higher-level" phrase boundaries (1,43–48). One approach to understanding the mechanism responsible for such effects is the π-gesture model of (1), in which it was hypothesized that boundary-related lengthening is caused by a special type of clock modulating system, a "π-gesture". This clock-modulating system, when active, slows down the rate of a hypothesized nervous system-internal global clock, relative to real time. Gestural activation dynamics evolve in the internal clock coordinate, and so gestural activation intervals are extended in time when a π-gesture is active. Furthermore, it was suggested in (1) that the degree of activation of a π-gesture varies in relation to the strengths of prosodic boundaries, such that stronger/higher-level boundaries are associated with greater π-gesture activation and hence more slowing.

How can the phenomenon of boundary-related lengthening be conceptualized in the current framework, where there is no global internal clock for gestural systems? A fairly straightforward solution is to recognize that in effect, each gestural system has its own "local clocks", in the form of the internal and external feedback TiRs, whose integration rates are modulated by self-attention. In that light, it is sensible to adapt the π-gesture mechanism by positing that self-attention effects on TiR parameters tend to be greater not only in the final set of gestures selected in each prosodic word (i.e. final μ-system), but also in the final set of co-selected conceptual systems (i.e. the final $\mathcal{C}$-system). As for why it is the final set of selected systems that induces these effects, we reason that speakers may attend to sensory feedback to a greater degree when there are fewer systems that remain to be selected. At the end of an utterance, there are no more systems that remain to be selected, and thus self-attention is greatest. We refer to this idea as the *selectional anticipation hypothesis*, because anticipation of upcoming selection events is proposed to distract a speaker from attention to feedback of their own speech. Although this hypothesis is admittedly a bit ad hoc, and alternative accounts should be considered, we show below that the implementation of this idea is sufficient to generate the lengthening that occurs at the ends of phrases.

Putting the above ideas together, Figure 12 shows how interval durations change as a function of attentional modulation. The utterance here is a competitively selected sequence of three syllables with forms CVC, CV, CVC, as shown in Figure 12A. Note that the organization of each syllable conforms to the hybrid control model, entailing that pre-vocalic timing is coordinative and vocalic/post-vocalic timing is feedback-based. As in Section 3.2, the integration rates of external (sensory) and internal TiRs, along with oscillator frequencies, are made to vary in response to changes in a control parameter λ; these relations are shown in Figure 12B. In addition, the integration rate parameters associated with the final set of gestures are even more strongly modulated by λ (dotted lines of Figure 12B), to implement the selectional anticipation hypothesis. The initiation times of gestures for each of the 11 values of λ that were simulated are shown vertically in Figure 12C.

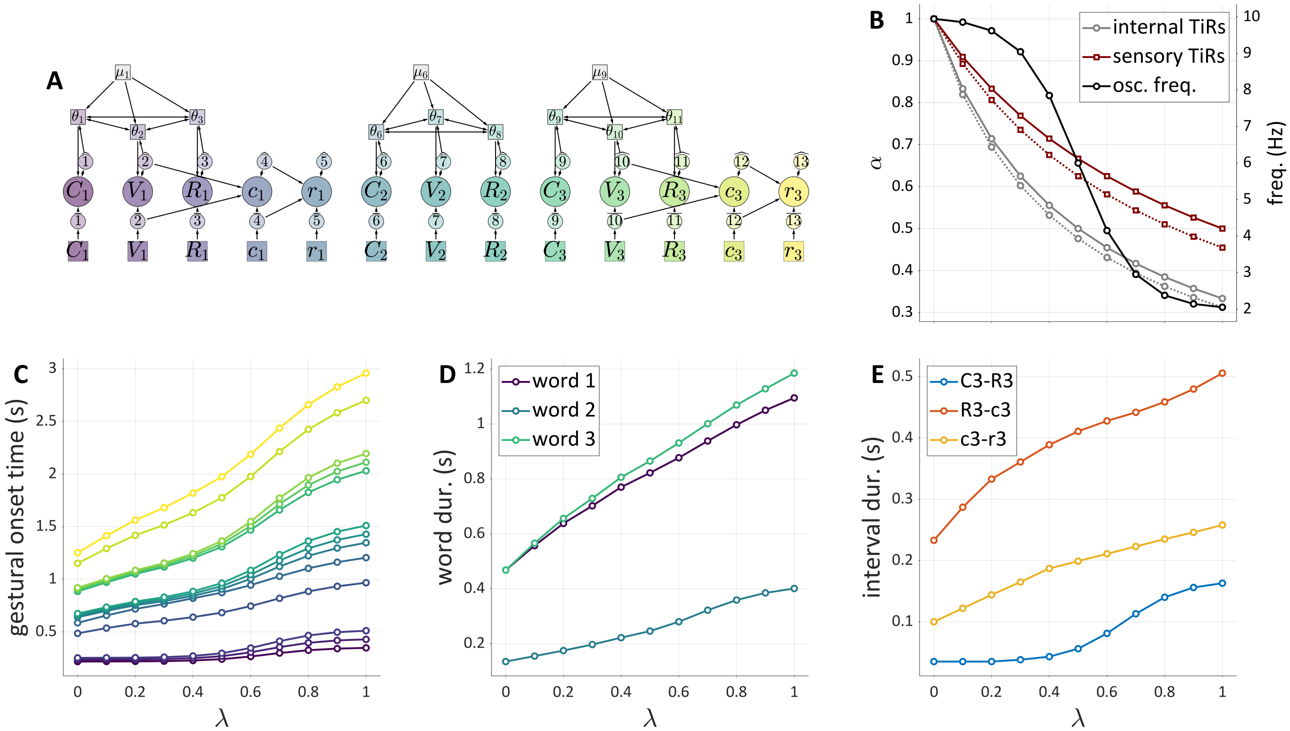

Figure 12. Simulation of variation in speech rate, as controlled by correlated changes in self-attention and pacing, both indexed by λ. (A) Model schema showing three syllables with the forms CVC, CV, and CVC. (B) Relations between λ and feedback TiR integration rates (α) and oscillator frequencies. (C) Times of gestural initiation for each value of λ simulated. (D, E) Word durations and interval durations of the third word.

By simulating variation in speech rate, we are able to generate some of the most essential predictions of the hybrid control model, introduced in Section 3.1. Recall that this model combined two hypotheses: prevocalic coordinative control and post-vocalic feedback-control. These hypotheses are associated with the following three predictions:

(i) *Prevocalic attenuation*. The prevocalic coordinative control hypothesis holds that initiation of the prevocalic constriction and release gestures, along with initiation of the vocalic gesture, is controlled by a system of coupled oscillators. Moreover, the frequency constraint hypothesis was shown in Section 3.2 to predict that intervals between these initiations attenuate as rate is increased or decreased. This effect can be seen in Figure 12E for the C_3-R_3 interval, which is the interval between constriction formation and release. In other words, the prediction is that prevocalic timing is only so compressible/expandible, no matter how quickly or slowly a speaker might choose to speak.

(ii) *Postvocalic expandability*. Conversely, the post-vocalic feedback-control hypothesis holds that there is a transition from internally to externally governed control, and that there should be no limits on the extent to which increasing self-attention can increase the corresponding interval durations. This prediction is shown in Figure 12E for the R_3-c_3 interval (which loosely corresponds to acoustic vowel duration) and the c_3-r_3 interval (related to constriction duration). These intervals continue to increase as attention to feedback is increased.

(iii) *Sensitivity to feedback perturbation*. Finally, a third prediction of the model is that, when external feedback governs post-vocalic control (as is predicted for slow rates), perturbations of sensory feedback will influence post-vocalic control but not prevocalic control.

How do these predictions fare in light of current evidence? The ideal tests of predictions (i) and (ii) require measurements of temporal intervals produced over a wide range of variation in global speech rate. Unfortunately, most studies of the effects of speech rate do not sufficiently probe extremal rates, since many studies use categorical adverbial instructions (e.g. *speak fast* vs. *speak normally* vs. *speak slowly*). One exception is a recent study using an elicitation paradigm in which the motion rate of a visual stimulus iconically cued variation in speech rate (49). Utterance targets were words with either intervocalic singleton or geminate bilabial nasals (/ima/ and /imma/). The study observed that the timing of constriction formation and release of singleton /m/ exhibited a nonlinear plateau at slow rates, similar to the prediction for the c_3-r_3 interval in Figure 12E. This is expected given the assumption that the formation and release gestures are organized in onset of the second syllable of the target words. In contrast, the constriction formation-to-release intervals of geminate /mm/ did not attenuate: they continued to increase in duration as rate slowed. This is expected if the initiation of the geminate bilabial closure is associated with the first syllable and its release with the second. Although the dissociation of effects of rate on singletons vs. geminates is not the most direct test of the hybrid model hypothesis, it shows that more direct tests are warranted.

Regarding prediction (iii), a recent study has indeed found evidence that post-vocalic intervals respond to temporal perturbations of feedback and that pre-vocalic intervals do not (50). This study found that subtle temporal delays of feedback imposed during a complex onset did not induce compensatory timing adjustments, while the same perturbations applied during a complex coda did. This dissociation in feedback sensitivity is a basic prediction of the hybrid model. Another recent study (51) has found that temporal perturbations induced compensatory adjustments of vowel duration but not of onset consonant duration (codas were not examined). There may be other reasons why temporal feedback perturbations have differential effects on prevocalic and vocalic/post-vocalic intervals, and certainly there is much more to explore with this promising experimental paradigm. Nonetheless, effects that have been observed so far are remarkably consistent with the predictions of the hybrid control model.

4 General discussion

The informal logic developed here has many consequences for phonological theories. Below we discuss three of the most important ones. First, the framework does not allow for direct control over the timing of articulatory target achievement, and we will argue that this is both conceptually desirable and empirically justified. Second, structural entities such as syllables and moras can be re-interpreted in relation to differences in the organization of control. Third, there is no need to posit the existence of different types of phrases, nor a hierarchical organization of phrases: the appearance of prosodic "structure" above the phonological word can reinterpreted more simply as variation in self-attention conditioned on selection of prosodic words.

4.1 No direct control of target achievement

Some researchers in the TD/AP framework have explicitly hypothesized that control of timing of target achievement is a basic function available in speech (52), or have implicitly assumed such control to be available (53). More generally, outside of the AP/TD framework, it has been argued that speakers prioritize control of the timing of articulatory and acoustic target events over control of the initiation of very same actions that are responsible for achieving those targets (48,54,55). "Target achievement" is defined here as a event in which the state of the vocal tract reaches a putative target state that is associated with a gestural system.

Direct control of the timing of gestural target achievement is prohibited by our logic because TiRs control when gestural systems become active and cease to be active, and neither of these events fully determines the time at which targets are achieved. The TiR framework of course allows for *indirect* control of target achievement timing, via the trivial fact that target achievement depends in part on when a gesture is activated. Yet other factors, which are outside the scope of the TiR model, play a role as well. In standard Task Dynamics (7) these factors include the strengths of the forces that gestural systems exert on a tract variable systems—both driving forces and dissipative damping forces—as well as how these forces are blended when multiple gestural systems are active. Or, in an alternative model of how gestures influence tract variable control systems (41), the relevant factors are the strengths, timecourses, and distributions of inhibitory and excitatory forces that gestural systems exert on spatial fields that encode targets. In either case, target achievement cannot be understood to be controlled directly by TiRs.

A major conceptual issue with direct control of target achievement is that it requires an unrealistically omniscient system which also has accurate knowledge of the future. In order to control exactly when a target is achieved, a control system must initiate a movement at precisely the right time, which in turn requires that the system is able to anticipate the combined influences on the vocal tract state of all currently active subsystems and all subsystems which might become active in the near future. This all-knowing planner must accomplish these calculations before the critical time at which the movement must be initiated. While such calculations are not in principle impossible, they do require a system which has access to an implausibly high degree of information from many subsystems.

A primary empirical argument for direct control of target achievement is premised on the claim that there is less variability associated with timing of target achievement than variability associated with timing of movement onsets. This is argued in (48,54) to suggest that timing of target achievement is not only independently controlled, but also prioritized over timing of movement initiation. The difference in variability upon which the argument is premised has been observed in non-speech studies in which an actor must hit or catch a moving object. Yet these sorts of non-speech examples do not necessarily translate to speech, because in articulation there are no uncontrolled moving objects that the effectors must collide with at the right place in space and time—speech is simply not like catching a ball. Indeed, only one study of speech appears to have concluded that there is less variability in target vs. initiation timing (56), and this interpretation of the data is highly questionable due to differences in how the two events were measured.

Empirically observed phonetic and phonological patterns indeed provide the strongest argument *against* direct control of target achievement timing. Phonetic reduction of targets, which can arise from insufficient allotment of time for a target to be achieved, is rampant in speech. The "perfect memory" example of (8) shows how at fast speech rates the word-final [t] can be not only acoustically absent but also quite reduced kinematically when the preceding and following velar and bilabial closures overlap. If speakers prioritized the timing of the [t] target relative to either the preceding or following targets, this sort of reduction presumably would happen far less often. The prevalence of historical sound changes which appear to involve deletion of constriction targets, argues against the notion that speakers are all that concerned with achieving targets. Certainly, the consequences of failing to achieve a target are usually not so severe: in order to recognize the intentions of speakers, listeners can use contextual information and acoustic cues that not directly related to target achievement. Rather than being a priority, our informal logic views target achievement as an indirect and often not-so-necessary consequence of activating gestural systems.

4.2 Reinterpretation of syllabic and moraic structure

Many phonological theories make use of certain structural entities—syllables (σ) and moras (μ)—as explanatory structures for phonological patterns. These entities are viewed as groupings of segments, with moras being subconstituents of syllables, as was shown in Figure 1B. Selection-coordination theory (3,4) has argued that these entities, rather than being parts of a structure, should be thought of as different classes of phonological patterns that are learned in different stages of a particular developmental sequence, over which the organization of control changes. This idea is referred to as the *holographic hypothesis*, because it holds that what appears to be a multi-level structure of syllables and moras is in fact a projection over developmental time of two single-level structures which do not exist simultaneously. This is loosely analogous to a hologram, which encodes a three-dimensional image in two dimensions.

The holographic hypothesis is exemplified in Figure 13 for a CVC syllable. Early in development, the post-vocalic constriction gesture is controlled entirely by sensory feedback (i.e., extra-gestural TiRs), and so phonological patterns learned at this time are associated with a moraic structure, reflecting a stronger differentiation in control of pre-vocalic and post-vocalic articulation. Subsequently, speakers learn to activate and deactivate the post-vocalic constriction/release with internal TiRs, process called *internalization*. This leads to initiation of the post-vocalic constriction before termination of the vocalic gesture, hence an increase in articulatory overlap/coarticulation. Phonological patterns learned in conjunction with this internalized organization of control are associated with syllables, rather than moras. Similar reasoning applies to other syllable shapes such as {C}{CV}→{CCV} and {CV}{V}→{CVV}, where developmental transitions in the internalization of control can account for cross-linguistic phonetic and phonological variation (3).

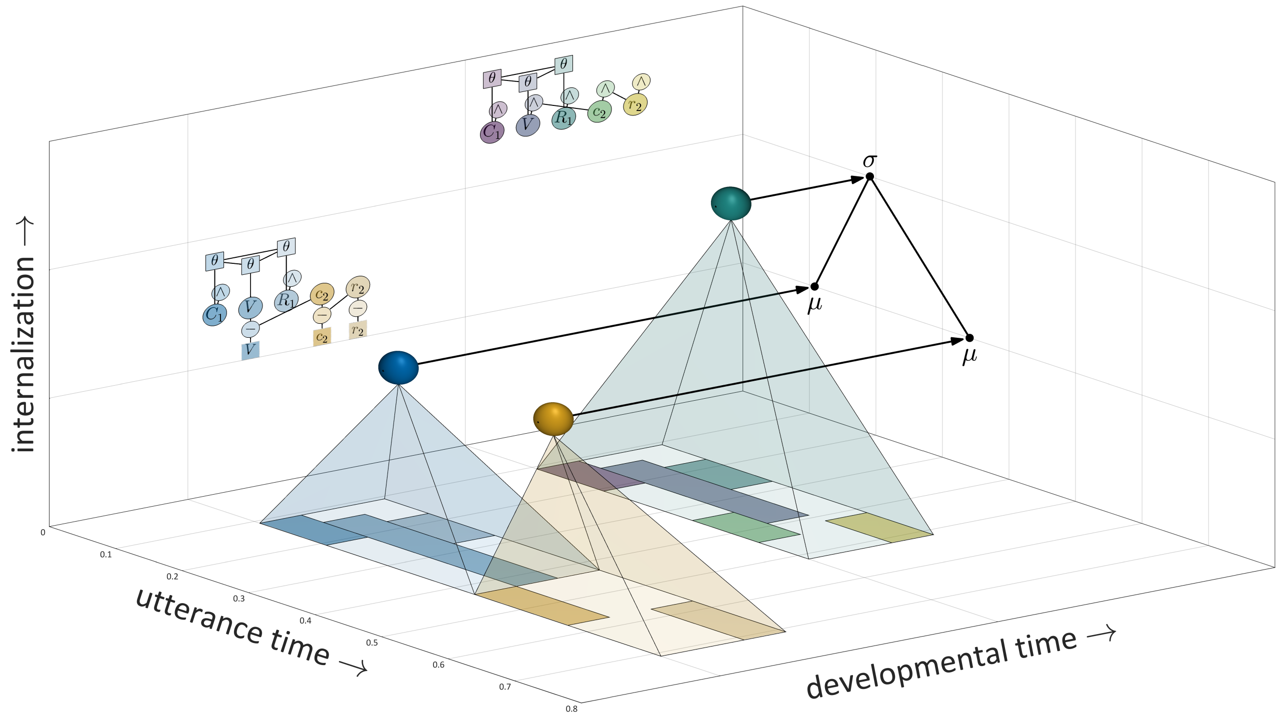

Figure 13. Visualization of the holographic hypothesis, for a CVC form. In an early stage of development, control over the post-vocalic constriction is based entirely on sensory feedback. Phonological patterns learned in this stage of development are described with moraic structure. In a later stage of development, control has been internalized, and phonological patterns learned in this stage are described with syllabic structure.

Exactly what causes internalization and governs its progression are open questions that presumably relate to information transmission. More internalization is associated with a greater rate of information production in speech, or in other words, increased efficiency of communication. Conversely, too much internalization can result in degrees of articulatory overlap which sacrifice perceptual recoverability (57–60), reflecting constraints on channel capacity. It is far from clear how these opposing considerations—information rate vs. channel capacity—might be mechanistically manifested in a model of utterance-timescale processes. Informational aspects of speech, which by definition require analysis of the space of possible state trajectories of gestural systems, necessarily involve attention to patterns on lifespan timescales and speech-community spatial scales. Thus the challenge lies in understanding how these relatively large timescale informational forces translate to changes in utterance-scale control.

4.3 Reinterpretation of prosodic phrase structure and boundaries

There are many prosodic theories in which prosodic words (ω) are understood to be hierarchically structured into various types of phrases. A "phrase" in this context simply refers to a grouping of prosodic words. Different types of phrases have been proposed, with two of the most popular being the "intonational phrase" (IP) and "intermediate phrase" (iP) from (61); these were shown in Figure 1B. Many theories additionally posit that these types of phrases can be recursively hierarchically structured (62–64), such that a given type of phrase can contain instances of itself. In general, the motivations for positing phrase structures of this sort are diverse and too complex to address in detail here, but most of them relate either to the likelihood that certain phonological patterns will occur in some portion of an utterance or to statistical patterns in measures of pitch or duration observed in longer utterances.

To provide an example, consider the question: *Who was in the library?*, answered with the utterance *Al and Bo or Cam were there*. This utterance has two probable interpretations, and in many theories these would be disambiguated by the prosodic structures shown in Figure 14 (A vs. B):

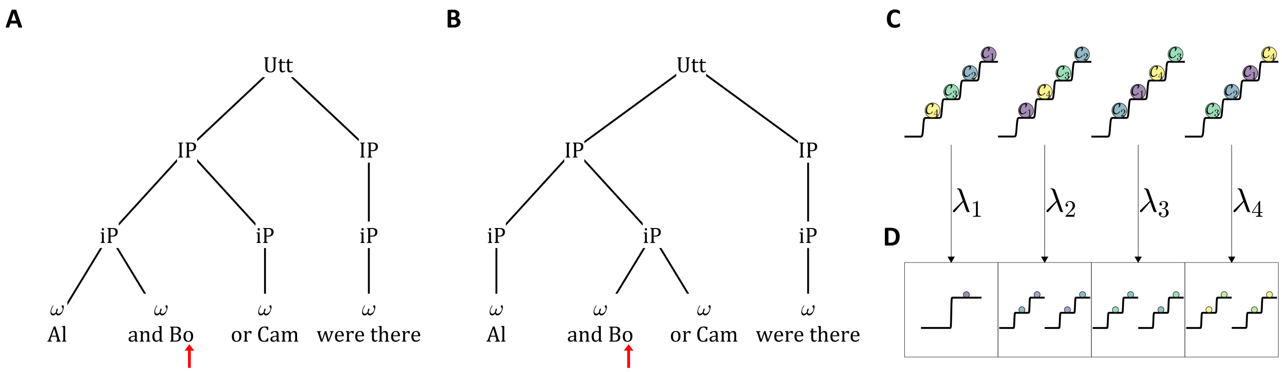

Figure 14. Hierarchical prosodic structure reinterpreted as variation in attentional modulation of control parameters. (A vs. B): alternative hierarchical prosodic structures purported to encode a difference in conceptual grouping. Red arrows indicate timepoint discussed in the text. (C, D) In different epochs of concept system selection, self-attention (λ) may differ, resulting in differences in temporal control.

The motivation for positing the structural distinction between (A) and (B) is that it can account for certain empirical patterns related to conceptual grouping. Consider specifically the period of time in the vicinity of the red arrows, near the end of the production of *Bo*, which is often conceptualized as a phrase "boundary". Here utterance (A), compared to (B), will tend to exhibit a larger fall of pitch, greater boundary-related lengthening, and a greater likelihood of a pause. The pitch of the following word may also start at a higher value. Hierarchical structural analyses hold that these differences occur because there is a "higher-level boundary" here in (A) than in (B), that is, an intermediate phrase boundary vs. a prosodic word boundary.

The logic of multilevel competitive selection makes hierarchical or recursive phrasal structure unnecessary. If anything, our framework corresponds to a flat, anarchical organization of prosodic words—though more appropriately it rejects the notion that prosodic words are parts of structures in the first place, and "boundaries" are seen as wholly metaphoric. How can regularities in intonational patterns such as in Figure 14 (A vs. B) be understood, without the notions of phrase hierarchies and boundaries?

Recall that each prosodic word is one set of co-selected concept systems, which are associated with some number of sets of co-selected gestural systems (Figure 11). Furthermore, recall that boundary-related lengthening was interpreted as a decrease in integration rates of feedback TiRs, and this parameter modulation is proposed to be greater for the last set of systems in a competitively selected set (the selectional anticipation hypothesis), as simulated in Figure 12. This reasoning leads to an alternative understanding of why there exists phonetic and phonological variation that correlates with prosodic organization: rather than being due to "structural" differences, the variation arises from differences in how TiR parameters are modulated for each prosodic word, as suggested by the arrows in Figure 14 (C and D). Rather than constructing a structure of prosodic words for each utterance, speakers simply learn to adjust self-attention in a way that can reflect conceptual relations between systems of concepts. Presumably many forms of discourse-related and paralinguistic information can be signaled in this way, including focus phenomena such as emphatic and contrastive focus.

5 Conclusion

To conclude, we return to the initial questions of this paper: (i) what determines the duration of that *shush* that you gave to the loud person in the library, and (ii) how do you slow down the rant to your friend in the coffee shop? According to the feedback-based logic of temporal control, your *shush* duration is most likely determined by a sensory feedback-based control system (an external, non-autonomous TiR), and depending upon various factors (how angry you are, how far away the loud student is), you will diminish the integration rate of the TiR and/or increase its threshold to extend the duration of the sound. Later on in the coffee shop, you slow down your rant in effect by doing the same thing: increasing self-attention.

There are several important conceptual and theoretical implications of our informal logic. First, all control of timing must be understood in terms of systems and their interactions, and this understanding involves the formulation of change rules to describe how system states evolve in time. Second, the systems which control timing do not "represent" time in any direct sense; the states of systems are defined in units of activation, and activation is never a direct reflection of elapsed time. Instead, it is more appropriate to say that timing is controlled via the integration of force, in combination with thresholds that determine when systems act. Third, the timing of target achievement is not a controlled event. Finally, much of the theoretical vocabulary that spans the range of timescales portrayed in Figure 1 is contestable, and new interpretations of empirical patterns can be derived from our logic. This applies to units such as syllables and moras, and also to hierarchical and recursive organizations of phrases. Ultimately the logic is useful because it facilitates a unified understanding of temporal patterns in speech, from the short timescale of articulatory timing to the large timescale of variation in speech rate.

6 Acknowledgments

I would like to thank members of the Cornell Phonetics Lab for discussion of the ideas in this manuscript.

7 Data Availability Statement

The code for running all simulations and generating all figures in this manuscript can be found on Github here: https://github.com/tilsen/TiR-model.git.

8 References

1. Byrd D, Saltzman E. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. Journal of Phonetics. 2003;31(2):149–80.

2. Tilsen S. A Dynamical Model of Hierarchical Selection and Coordination in Speech Planning. PloS one. 2013;8(4):e62800.

3. Tilsen S. Selection and coordination: The articulatory basis for the emergence of phonological structure. Journal of Phonetics. 2016;55:53–77.

4. Tilsen S. Selection-coordination theory. Cornell Working Papers in Phonetics and Phonology, 2014. 2014;24–72.

5. Tilsen S. Three mechanisms for modeling articulation: selection, coordination, and intention. 2018. (Cornell Working Papers in Phonetics and Phonology 2018).

6. Kelso JA, Saltzman EL, Tuller B. The dynamical perspective on speech production: Data and theory. Journal of Phonetics. 1986;14(1):29–59.

7. Saltzman E, Munhall K. A dynamical approach to gestural patterning in speech production. Ecological Psychology. 1989;1(4):333–82.

8. Browman C, Goldstein L. Tiers in articulatory phonology, with some implications for casual speech. Between the grammar and physics of speech. 1990;341–76.

9. Ladefoged P. Vowels and consonants: An introduction to the sounds of the world. Blackwell Publications; 2001.

10. Port RF, Leary AP. Against formal phonology. Language. 2005;927–64.

11. Jordan MI. Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986. California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science; 1986.

12. Browman C, Goldstein L. Some notes on syllable structure in articulatory phonology. Phonetica. 1988;45(2–4):140–55.

13. Browman C, Goldstein L. Articulatory phonology: An overview. Phonetica. 1992;49(3–4):155–80.

14. Kelso JAS, Tuller B. Intrinsic time in speech production: theory, methodology, and preliminary observations. Sensory and motor processes in language Hillsdale, NJ: Erlbaum. 1987;203:222.

15. Saltzman E, Nam H, Krivokapic J, Goldstein L. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In: Proceedings of the 4th international conference on speech prosody. Brazil: Campinas; 2008. p. 175–84.

16. Goldstein L, Byrd D, Saltzman E. The role of vocal tract gestural action units in understanding the evolution of phonology. In: Action to language via the mirror neuron system. Cambridge: Cambridge University Press; 2006. p. 215–49.

17. Schöner G. Timing, clocks, and dynamical systems. Brain and cognition. 2002;48(1):31–51.

18. Sorensen T, Gafos A. The gesture as an autonomous nonlinear dynamical system. Ecological Psychology. 2016;28(4):188–215.

19. Browman C, Goldstein L. Gestural syllable position effects in American English. Producing speech: Contemporary issues. 1995;19–33.

20. Tilsen S. Exertive modulation of speech and articulatory phasing. Journal of Phonetics. 2017;64:34–50.

21. Tilsen S, Zec D, Bjorndahl C, Butler B, L'Esperance M-J, Fisher A, et al. A cross-linguistic investigation of articulatory coordination in word-initial consonant clusters. Cornell Working Papers in Phonetics and Phonology. 2012;2012:51–81.

22. Marin S, Pouplier M. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. Motor Control. 2010;14(3):380–407.

23. Buzsáki G, Draguhn A. Neuronal oscillations in cortical networks. science. 2004;304(5679):1926–9.

24. Buzsaki G. Rhythms of the Brain. Oxford University Press; 2006.

25. Nam H. Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. In: In J Cole & J I Hualde (Eds), Laboratory phonology. Berlin, New York: Walter de Gruyter; 2007. p. 483–506.

26. Tilsen S. Effects of syllable stress on articulatory planning observed in a stop-signal experiment. Journal of Phonetics. 2011;(39):642–59.

27. Cai S, Ghosh SS, Guenther FH, Perkell JS. Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. The Journal of Neuroscience. 2011;31(45):16483–90.

28. Houde JF, Jordan MI. Sensorimotor Adaptation in Speech Production. Science. 1998 Feb 20;279(5354):1213–6.

29. Larson CR, Burnett TA, Bauer JJ, Kiran S, Hain TC. Comparison of voice f responses to pitch-shift onset and offset conditions. The Journal of the Acoustical Society of America. 2001;110:2845.

30. Purcell DW, Munhall KG. Compensation following real-time manipulation of formants in isolated vowels. The Journal of the Acoustical Society of America. 2006;119(4):2288–97.

31. Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. Neuroimage. 2008;39(3):1429–43.

32. Villacorta VM, Perkell JS, Guenther FH. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. The Journal of the Acoustical Society of America. 2007;122(4):2306–19.

33. Grossberg S. The adaptive self-organization of serial order in behavior: Speech, language, and motor control. Advances in Psychology. 1987;43:313–400.

34. Bullock D. Adaptive neural models of queuing and timing in fluent action. Trends in cognitive sciences. 2004;8(9):426–33.

35. Bullock D, Rhodes B. Competitive queuing for planning and serial performance. CAS/CNS Technical Report Series. 2002;3(003):1–9.

36. Bohland JW, Bullock D, Guenther FH. Neural representations and mechanisms for the performance of simple speech sequences. Journal of cognitive neuroscience. 2010;22(7):1504–29.

37. Glasspool DW. Competitive queuing and the articulatory loop. Psychology Press; 2014.

38. Bhutani N, Sureshbabu R, Farooqui AA, Behari M, Goyal V, Murthy A. Queuing of concurrent movement plans by basal ganglia. Journal of Neuroscience. 2013;33(24):9985–97.

39. Kristan WB. Behavioral sequencing: Competitive queuing in the Fly CNS. Current Biology. 2014;24(16):R743–6.

40. Tilsen S. Syntax with oscillators and energy levels. Berlin: Language Science Press; 2019. (Studies in Laboratory Phonology).

41. Tilsen S. Motoric mechanisms for the emergence of non-local phonological patterns. Frontiers in Psychology. 2019;10:2143.

42. Tilsen S. Space and time in models of speech rhythm. Annals of the New York Academy of Sciences. 2019;1453(1):47–66.

43. Byrd D, Saltzman E. Intragestural dynamics of multiple prosodic boundaries. Journal of Phonetics. 1998;26:173–99.

44. Byrd D. Articulatory vowel lengthening and coordination at phrasal junctures. Phonetica. 2000;57(1):3–16.

45. Krivokapić J. Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2014;369(1658):20130397.

46. Byrd D, Krivokapić J, Lee S. How far, how long: On the temporal scope of prosodic boundary effects. J Acoust Soc Am. 2006 Sep;120(3):1589–99.

47. Turk AE, Shattuck-Hufnagel S. Multiple targets of phrase-final lengthening in American English words. Journal of Phonetics. 2007;35(4):445–72.

48. Turk A, Shattuck-Hufnagel S. Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production. Frontiers in psychology. 2020;10:2952.

49. Tilsen S, Hermes A. Nonlinear effects of speech rate on articulatory timing in singletons and geminates. In: Proceedings of the 12th International Seminar on Speech Production. 2020.

50. Oschkinat M, Hoole P. Compensation to real-time temporal auditory feedback perturbation depends on syllable position. The Journal of the Acoustical Society of America. 2020;148(3):1478–95.

51. Karlin R, Naber C, Parrell B. Auditory Feedback Is Used for Adaptation and Compensation in Speech Timing. Journal of Speech, Language, and Hearing Research. 2021;64(9):3361–81.

52. Gafos AI. A grammar of gestural coordination. Natural Language & Linguistic Theory. 2002;20(2):269–337.

53. Shaw J, Gafos AI, Hoole P, Zeroual C. Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. Phonology. 2011;28(03):455–90.

54. Turk A, Shattuck-Hufnagel S. Timing in talking: what is it used for, and how is it controlled? Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1658):20130395.

55. Turk A, Shattuck-Hufnagel S. Speech Timing: Implications for Theories of Phonology, Speech Production, and Speech Motor Control. Vol. 5. Oxford University Press, USA; 2020.

56. Perkell JS, Matthies ML. Temporal measures of anticipatory labial coarticulation for the vowel/u: Within-and cross-subject variability. The Journal of the Acoustical Society of America. 1992;91(5):2911–25.

57. Chitoran I, Goldstein L. Testing the phonological status of perceptual recoverability: Articulatory evidence from Georgian. In: Proc of the 10th Conference on Laboratory Phonology, Paris, June 29th–July 1st. 2006. p. 69–70.

58. Gick B, Campbell F, Oh S, Tamburri-Watt L. Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. Journal of Phonetics. 2006;34(1):49–72.

59. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological review. 1967;74(6):431.

60. Fowler CA, Rosenblum LD. The perception of phonetic gestures. Modularity and the motor theory of speech perception. 1991;33–59.

61. Beckman ME, Pierrehumbert JB. Intonational structure in Japanese and English. Phonology. 1986;3:255–309.

62. Féry C. Recursion in prosodic structure. Phonological Studies. 2010;13:51–60.

63. Ladd DR. Intonational phrasing: the case for recursive prosodic structure. Phonology. 1986;3:311–40.

64. Ito J, Mester A. Prosodic subcategories in Japanese. Lingua. 2013;124:20–40.