Psychological Review

© 2019 American Psychological Association
ISSN: 0033-295X

2020, Vol. 127, No. 2, 281–304
http://dx.doi.org/10.1037/rev0000172

Control of Speaking Rate Is Achieved by Switching Between Qualitatively Distinct Cognitive "Gaits": Evidence From Simulation

Joe Rodd and Hans Rutger Bosker
Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, and Radboud University

Mirjam Ernestus
Radboud University and Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

Phillip M. Alday
Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

Antje S. Meyer and Louis ten Bosch
Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, and Radboud University

That speakers can vary their speaking rate is evident, but how they accomplish this has hardly been studied. Consider this analogy: When walking, speed can be continuously increased, within limits, but to speed up further, humans must run. Are there multiple qualitatively distinct speech "gaits" that resemble walking and running? Or is control achieved by continuous modulation of a single gait? This study investigates these possibilities through simulations of a new connectionist computational model of the cognitive process of speech production, EPONA, that borrows from Dell, Burger, and Svec's (1997) model. The model has parameters that can be adjusted to fit the temporal characteristics of speech at different speaking rates. We trained the model on a corpus of disyllabic Dutch words produced at different speaking rates. During training, different clusters of parameter values (regimes) were identified for different speaking rates. In a 1-gait system, the regimes used to achieve fast and slow speech are qualitatively similar, but quantitatively different. In a multiple gait system, there is no linear relationship between the parameter settings associated with each gait, resulting in an abrupt shift in parameter values to move from speaking slowly to speaking fast. After training, the model achieved good fits in all three speaking rates. The parameter settings associated with each speaking rate were not linearly related, suggesting the presence of cognitive gaits. Thus, we provide the first computationally explicit account of the ability to modulate the speech production system to achieve different speaking styles.

Keywords: speech production, speech rate, connectionist models, executive control, lexical access

Supplemental materials: http://dx.doi.org/10.1037/rev0000172.supp

Speaking is a uniquely human behavior. It is by nature temporal: concepts and ideas are encoded as a stream of rapidly fluctuating sound, and the correct ordering and duration of the components is of crucial importance for intelligibility and conveying meaning. At the same time, there is great variability in the timing of speech sounds: different speakers have different habitual speech rates, and

This article was published Online First December 30, 2019.

Joe Rodd, Psychology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, and Centre for Language Studies, Radboud University; Hans Rutger Bosker, Psychology of Language Department, Max Planck Institute for Psycholinguistics, and Donders Institute for Brain and Behaviour, Radboud University; Mirjam Ernestus, Centre for Language Studies, Radboud University, and Max Planck Institute for Psycholinguistics; Phillip M. Alday, Psychology of Language Department, Max Planck Institute for Psycholinguistics; Antje S. Meyer, Psychology of Language Department, Max Planck Institute for Psycholinguistics, and Donders Institute for Brain and Behaviour, Radboud University; Louis ten Bosch, Max Planck Institute for Psycholinguistics, and Centre for Language Studies, Radboud University.

This research was supported by Netherlands Organization for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium. We are grateful to Jeroen van Paridon, Laurel Brehm, and Lou Boves for useful comments on our modelling and analysis methods. Elements of the work presented in the article have previously been presented at the following venues: seminars at the University of Munich (January 2018), University of Düsseldorf (March 2019), University of Amsterdam (April 2019), and University of Utrecht (April 2019); and at the following workshops and conferences: "International Workshop on Speech Production" (Nijmegen, June 2018); "Architectures and Mechanisms for Language Processing" (AMLaP, Berlin, September 2018); "Crossing the Boundaries: Language in Interaction" (Nijmegen, April 2019); "Phonetics and Phonology in Europe" (PaPE Lecce, June 2019).

Correspondence concerning this article should be addressed to Joe Rodd, Psychology of Language Department, Max Planck Institute for Psycholinguistics, Postbus 310 6500 AH Nijmegen, the Netherlands. E-mail: joe.rodd@mpi.nl

individual speakers can vary their speech rate from situation to situation, and even within utterances in the same conversation (e.g., Miller, Grosjean, & Lomanto, 1984; Quené, 2008). A portion of this variation presumably arises to accommodate different communicative situations: speakers may slow down to provide listeners with sufficient time to extract the necessary details from the acoustic signal (e.g., Bosker & Cooke, 2018; Cooke, King, Garnier, & Aubanel, 2014; Lindblom, 1990). Alternatively, they may speed up, for instance to convey more content in the same period of time. Listeners use speech rate information in shaping their perception (Dilley & Pitt, 2010; Kaufeld, Ravenschlag, Meyer, Martin, & Bosker, 2019; Maslowski, Meyer, & Bosker, 2019), making control of speech rate an essential communicative skill.

The fact that humans have control over the rate at which they speak means that they are capable of adjusting the cognitive apparatus that plans speech, from the selection of words to the tightly coordinated movements of the articulators of the vocal tract. Understanding how speech planning is controlled can give us insights into how the apparatus itself works. Given the large degree of speaker-controlled variability in speech, identifying the mechanisms of control over speech planning is also important in its own right. In the present study, we examine the control processes speakers may engage to achieve different speech rates.

Speech production is classically characterized as a modular, feed-forward processing system (e.g., Dell & O'Seaghdha, 1992; Levelt, 1989; Levelt, Roelofs, & Meyer, 1999; Stemberger, 1985). After a meaning representation has been selected ("conceptualization"), the lexical selection stage begins, where abstract representations of words that best correspond to the conceptual message are selected. Processes of word form encoding then construct detailed word form representations. These stages together can be considered as a formulation phase. Once a word form representation is selected, a motor execution phase is entered, where movement commands for the articulatory apparatus (e.g., the tongue, lips, vocal chords) are calculated, carried out, and monitored (Guenther, 2016; Tourville & Guenther, 2011). Because speakers typically plan as late as possible, rather than storing a preplanned utterance in working memory (e.g., Damian & Dumay, 2007; Kello, Plaut, & MacWhinney, 2000; Levelt, 1989; Levelt et al., 1999), the formulation system must keep up with the desired rate of articulation, requiring modulation of its operation to maintain synchronization.

"Gaits" in Speech Production

In a working model of the production system with formulation and execution phases, adjustment in speaking rate results from adjusting the state of the formulation system; to speak slowly we shift to a regime that results in slow speech and to speak fast we shift to a regime that causes speech to emerge more quickly. How are these regimes related to each other? How does the regime invoked to produce slow speech differ from the regime invoked to produce medium rate speech?

The control mechanisms engaged to regulate speaking rate at the level of utterance planning and preparation are largely unknown. A more concrete and readily observable system that operates at a continuously varying range of speeds is that of human and animal locomotion. In humans, walking and running gaits are adopted to achieve movement at different speeds. The movement patterns of walking and running are qualitatively different; in walking, at least one foot is on the ground at all times, while in running, both feet are raised from the ground simultaneously for part of the cycle (Alexander, 1989; Minetti, 1998). A continuous range of movement speeds can be achieved by first increasing the speed of walking, and then switching to a running gait to speed up further. Alongside hard limits on feasibility of certain gaits at certain speeds, the selection of locomotive gaits is tightly linked to their relative efficiency. In horses, which typically have walking, trotting, and galloping gaits, each gait has a clear "sweet spot" speed, at the approximate center of the range of speeds achievable with that gait, where exertion (ml O_2 consumed to move 1 m) is minimized (Hoyt & Taylor, 1981, their Figure 2). Horses and migratory animals select these speeds preferentially (Pennycuick, 1975), and avoid the inefficient speeds in the shoulder of each gait. This feature of gaited systems previously inspired speech researchers working at the level of articulatory movements, who link qualitatively different mechanical realizations of speech movements to their relative efficiency to achieve a required standard of intelligibility (e.g., Pouplier, 2012), as predicted by the hyper- and hypoarticulation theory (Lindblom, 1990).

Pouplier (2012) related the metabolic equivalence of the optima of the locomotive gaits to speaking, conceptualizing the gaits of speech as equally optimal coordination modes, suitable for different contexts. This holds well for the execution phase of speech production, which incorporates motor planning and articulation, where there are "many roads to Rome": Different gestural coordination configurations, which are chosen between according to local context, can lead to acoustic outcomes that are equivalent for the listener. For instance, speakers can make use of alternative vocal tract configurations to achieve speech sounds when articulatory freedom is constrained (Lindblom, Lubker, & Gay, 1977). Immediately adjacent speech sounds also condition the selection of alternative articulatory configurations, so as to minimize the articulator movement required (e.g., Boyce & Espy-Wilson, 1997). This reconfiguration can be thought of as analogous to switching between gaits in locomotion.

More global contextual factors such as prosody and speech rate can also lead to gestural reconfiguration in the execution component, for instance in coda consonant resyllabification, whereby a consonant may be realized in a way more similar to an onset consonant (Scobbie & Pouplier, 2010) in rate-scaling experiments. Similarly, antiphase synchronization of gestures tends to reconfigure to in-phase synchronization as rate increases (Kelso, Saltzman, & Tuller, 1986); for instance in West Andalusian Spanish, Parrell (2012) finds that speakers shift from antiphase oral-glottal coordination in sequences like ['ka.^{h}ta] from /casta/, *"caste"* (with preaspiration before the [ti]) to in-phase coordination ['ka.t^{h}a], by making the tongue articulation of the /t/ earlier so it occurs at the same time as the glottal opening.

Alternatively, the speech planning apparatus might be purely linearly up- or down-regulated in response to changes in required speaking rate. This is the case for motor tasks where temporal precision is required, in both gross motor movements (Wright & Meyer, 1983), and fine movement requiring extensive coordination, such as piano playing (Bella & Palmer, 2011).

Approach Adopted in This Study

We extend the metaphor of gaitedness to the *psychological* system of speaking rate control. We ask whether there are multiple cognitive gaits in speech planning that resemble locomotive gates. Without a choice of gaits, the cognitive regimes adopted to achieve different speaking rates would be similar in nature, but only quantitatively different. In other words, the difference between the regimes required to produce slow and medium speech would be similar to the difference between the regimes required to produce medium and fast speech. This is akin to only having one gait, which can be sped up or slowed down linearly. Alternatively, with multiple gaits of speech planning, the regimes would differ from each other in a nonlinear way, with a qualitative difference between, for instance, the regimes adopted for slow rates (walk-speaking) and the regimes adopted for fast rates (run-speaking).

We address the question of how speakers control speech rate. More concretely, we aimed to ascertain how the cognitive regimes that are associated with each speaking rate relate to each other, to assess if multiple gaits might be present. To do this, we constructed a family of computationally implemented connectionist models of the formulation phase of speech planning (Strand 1), and explored how each model variant could be optimized to mimic the temporal properties of natural word productions taken from a speech corpus elicited at different cued speaking rates. We then evaluated the performance of the optimized model variants. This process allowed us to identify optimal model parameter settings associated with producing speech at a given rate, which provide a window onto the arrangement of the regimes of the underlying cognitive systems (Strand 2).

Computational Model (Strand 1)

A computational model of the speech planning system provides a psycholinguistic sandbox to explore how the regimes adopted to achieve speech at different speaking rates relate to each other. We propose such a computational model, EPONA. EPONA has parameters that determine its behavior (controlling features such as rate of activation spreading, rate of activation decay, and connection weightings). These parameters can be optimized to cause the model to optimally fit speech data produced at different speaking rates. The sets of parameter values chosen by the model for each rate condition mirror the regimes of the cognitive system that the model emulates. More concretely, we adopted an optimization procedure which identified the parameter values required to fit the distributions of three durational features measured from elicited productions of disyllabic words: first syllable durations, second syllable durations, and overlap durations. The distributions of these durational features together form a “fingerprint” of the regime of the speech production system engaged to achieve that speaking rate. This process was repeated for three different speaking rates: fast, medium, and slow.

The theoretical model that we selected as inspiration for EPONA is that of Dell, Burger, and Svec (1997). The model is a good starting point because it captures the ability to produce sequences of elements from a hierarchical structure. The model separates the encoding of the segmental content of the word from the encoding of the metrical structure (the ordering and timing of the segmental content, and suprasegmental content such as word stress). EPONA inherits this property.

How do Regimes Relate to Each Other? (Strand 2)

The parameters of the EPONA model can be thought to represent the regimes of the cognitive system that underlie natural speech production at different rates. The different regimes of the system exist as locations in a multidimensional “parameter space,” where the parameters form the dimensions.

With a sample of three speaking rates, and assuming that each rate is associated with a single regime, there are five logical possibilities for how the regimes might be arranged with respect to each other. (a) The cognitive system has a single gait, and different speaking rates are achieved by continuous adjustment of this single gait. This is akin to only walking, but walking at three different speeds. (b) The cognitive system has three gaits, one for each speaking rate. These three gaits are qualitatively different, like walking, trotting, and galloping in horse locomotion. The cognitive system has two gaits, grouping the medium speaking rate with either the slow rate (c) or the fast rate (d). Finally, (e) The cognitive system has two gaits, a habitual gait adopted for the medium speaking rate, and an exceptional gait adopted for slow and fast speaking rates. This fifth option supposes that there is a default gait for the most frequently used speed, and that a fall-back “all purpose” gait is adopted for other rates.

In the single-gait scenario, the three regimes would be arranged along a single axis in parameter space. In a multiple gait scenario, the three regimes would be arranged in a triangle in parameter space. Each side of the triangle is potentially the axis of a gait to which two regimes belong. For each axis, if both speaking rate regimes belong to the same gait, we would expect a continuous, linear variation in the predictions of models fitted at points along the axis. If, however, the two regimes belong to different gaits, we would expect to see a nonlinearity at some point along the axis, indicating a shift from the area of parameter space associated with one gait to the area of parameter space associated with the other. To distinguish between the single and multiple gait scenarios, we examined the results of the optimization procedure undertaken in Strand 1 to identify the arrangement of the regimes in parameter space. To distinguish between various two- and three-gait scenarios, we fitted additional models at points along the axes between the three regimes, and assessed the predicted “fingerprints” for (non)linearity by means of Bayesian statistical modeling.

Serial Order in Speech Production and the Dell et al. (1997) Model

The core task of the formulation process is to ensure that after a lexical concept becomes active at the conceptual-formulation frontier, the gestural scores required to produce it become active at the frontier between formulation and motor execution. In this article, we will follow Levelt, Roelofs, and Meyer (1999) and Tourville and Guenther (2011) in assuming that the gestural score representation encodes the relative onset and offset times of abstract gestures (comparable with the gestures described by e.g., Browman & Goldstein, 1992) of a single syllable, and that this representation is shared by formulation and motor execution to allow activation to spread. In the execution component, a more concrete motor plan and auditory and somatosensory expectations are retrieved for this gestural score (Guenther, 2016; Tourville & Guenther, 2011).

A naive connectionist model of this process might assume direct connections from each word node to the relevant syllable nodes. Asking such a model to predict the temporal organization of a multisyllabic word such as the Dutch word *snavel* /'snaː.vəl/ "beak," however, will fail: /'snaː/ and /vəl/ will become active simultaneously. A successful model therefore needs to account for serial order; the fact that sequences of speech sounds are overwhelmingly often produced in the correct order (one or two errors per 1,000 words; Garnham, Shillcock, Brown, Mill, & Cutler, 1981), despite the subunits of each word presumably being activated from a single word-level parent node.

It is not trivial to construct a model that, in response to activation in a single parent node, can activate and then deactivate child elements in a sequence in turn. In the speech production domain, the most prominent model to deal with serial ordering is that of Dell et al. (1997, hereafter the DBS model). Dell et al. (1997) enumerate the requirements of serial ordering: preparation of the future, activation of the present and suppression of the past. That is, an ideal model should (a) prime upcoming syllables, (b) activate them at the correct time, and (c) deactivate them once they have been produced.

An example instantiation of the EPONA model capable of producing three Dutch disyllabic words is illustrated in Figure 1. The word-level input "plan nodes" are shown at the top of the model. At the bottom of the model are the syllable-level gestural score "content nodes." In between, there are two top-down routes along which activation can flow. The first route connects the plan nodes directly to the content nodes (shown with dashed red lines in Figure 1). The connections of this route are responsible for encoding the segmental content of the word, so we term it the "segmental route." The second route is responsible for maintaining correct serial order of syllables and encoding the metrical structure of the words by means of a frame node, which represents the word-level metrical structure, so we term it the "metrical route." The concept of separating the planning of segmental content and metrical structure into separate streams and employing a frame to enforce serial order is well established in framed-based psycholinguistic models of the production system (Bock, 1982; Dell, 1986; Garrett, 1976; Levelt, 1989; MacKay, 1972; Shattuck-Hufnagel, 1979; Stemberger, 1991). Note that throughout this article, C indicates a consonant, V indicates a vowel, ' indicates the syllable with primary stress, while a period (.) indicates the syllable boundary.

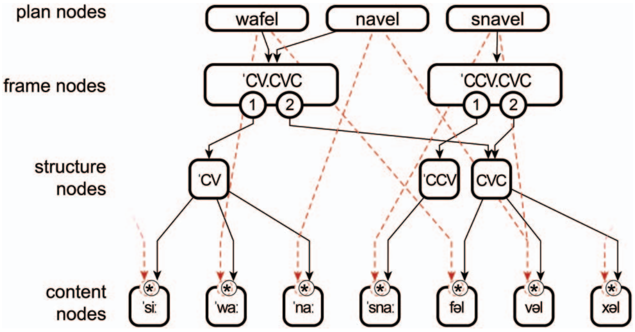

Figure 1. An instance of the EPONA model containing the nodes necessary to produce the Dutch disyllabic words *wafel* ['waː.fəl] "waffle," *navel* ['naː.vəl] "navel" and *snavel* ['snaː.vəl] "beak." The segmental route is shown with red dashed connections. At the top level, there is a unique plan node for each word. Frame nodes are shared between words with the same metrical structure (*wafel* and *navel* both have a 'CV.CVC structure, so are connected to the same frame node). Each frame node has multiple output ports (here numbered 1 and 2), one associated with each child element of the sequence. Each port is connected to a structure node. In turn, each structure node is connected to all content nodes representing syllables with the relevant metrical structure. Structure nodes and content nodes are also shared between words. Multiplication in the content nodes (represented by asterisks) ensures that only syllables receiving input from both routes become active. See the online article for the color version of this figure.

Frame-based models have two key advantages compared to models without them. First, because they separate information about sequential ordering from segmental information, they can explain the ordering of novel sequences without additional learning: if the correct frame and the correct content are known, previous separate experience with the frame and the content can be combined to produce the sequence correctly. Second, they account for the observation that errors where subelements are misordered within a sequence are overwhelmingly outnumbered by errors where elements from the same position in the sequence exchange ("caterpillar" → "patterkiller") or are copied between adjacent sequences. A model without frames would predict much more frequent misorderings of the elements within a sequence than is observed (Boomer & Laver, 1968; MacKay, 1970; Vousden, Brown, & Harley, 2000; Vousden & Maylor, 2006).

The metrical route is shown in Figure 1 with solid black arrows. Aside from the frame node, there are structure nodes, which are connected to all content nodes sharing a metrical structure at the syllable level. The first connection in the metrical route passes activation from the plan node to the relevant frame node. The frame node has an output port for each syllable in the word, so in our case, two ports. The first port is connected to a structure node for the metrical shape of the first syllable of the word. The second port is connected to a structure node for the metrical shape of the second syllable of the word. A mechanism within the frame node ensures the activation initially flows primarily from the first port, and subsequently from the second port; we will address the nature of this mechanism and the activation flows it generates shortly. The structure nodes therefore receive activation asynchronously: First the structure node representing the shape of the first syllable becomes active, and then the structure node representing the shape of the second syllable. The structure nodes spread their activation to all the content nodes that share that structure. In the content nodes, the incoming activation from the metrical route is multiplied by the incoming activation from the segmental route, meaning that nonzero activation must be received from both streams for the content node to become activated. The activation in the content nodes can be considered to be the output of the DBS model.

We will now turn to the frame node, which generates activation streams for each syllable in response to receiving activation from the word node above it. The DBS model is agnostic regarding the precise nature of the serial order mechanism employed in the frame node. Rather than including a pure-connectionist mechanism such as a competitive queue in the frame node (e.g., Hurlstone, Hitch, & Baddeley, 2014), Dell et al. (1997) construct a transparent model that exhibits serial-order behavior. This has the advantage of simplicity and interpretability.

In EPONA, the frame nodes directly produce parametrically defined activation patterns for each of the ports after they receive activation from the plan node. The ports can produce activation at three (parametrically defined) activation levels: baseline activation, partial activation, and full activation.

Activation is produced at these levels in a specific order (depicted in Figure 2). Before word onset, both ports produce baseline activation. The activation pattern for the first port (solid lines) begins with a period of partial activation, then a period of full activation, then baseline activation. The activation pattern for the second port (dashed) begins with baseline activation, then partial activation, then full activation, then baseline activation. The second port is therefore producing the same pattern as the first port, but delayed by the duration of one period. The partial activation level is proposed by Dell et al. (1997) as a means to prime the "future" (the next content to be produced). The full activation level is associated with activating the "present" (the content currently being produced). The baseline activation state serves as the baseline for ports connected to items that have not yet been produced, and is also associated with deactivating the "past" (content that has already been produced).

Mechanics of the Model

Dell et al. (1997) describe a mechanism that accounts for serial order behavior in speech production. They used the model to predict probabilities of speech errors. Error probabilities were calculated directly from predicted activation levels. To do so, it was not necessary to extract precise onset and offset times from the model. Rather than examining errors, we seek to understand how speakers adjust their speaking rate in correct utterances. To do so, we propose EPONA, a model that borrows its conception and underlying connectionist architecture from DBS. EPONA is able to predict the onset and offset times of syllable level planning units, and to model differences between speaking rates. EPONA differs from DBS in the specification of the timing behavior of the frame node, and extends it to add a rudimentary operationalization of the execution component. EPONA is implemented computationally, and is tested with speech timing data, rather than speech error proportions.

Timing in the Frame Node

The DBS model assumes that all the periods of the activation patterns associated with the ports of the frame node have equal duration. A model with this assumption is sufficient for the prediction of the rate of serial order errors, but it is improbable that such a model will be successful in fitting the relative onset and offset times of syllables in real speech, where the durations of syllables in a word are rarely equal (varying as a product of, among other things, the number of segments, the specific segments involved, the stress status of the syllable, and phonological processes such as final lengthening: Booij, 1995; Cambier-Langeveld, Nespor, & van Heuven, 1997; Slootweg, 1988). There are (at least) two ways that this constraint could be relaxed to allow the duration of full activation on each port to differ (and thus the overt production of each syllable), which should make the frame node more effective in encoding the metrical properties of the word shape it represents. These possibilities are described in the remainder of this section. We construct variants of EPONA consistent with each possibility.

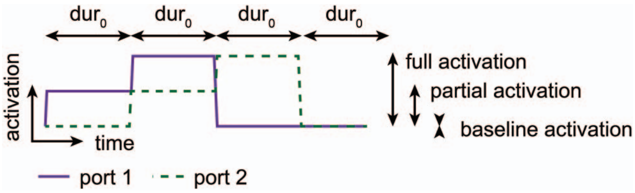

Figure 2. The activation patterns produced by the frame node for Port 1 (purple, solid) and Port 2 (green, dashed). See the online article for the color version of this figure.

The present implementation of EPONA produces only disyllabic words, but the mechanisms described here could be adapted to produce more syllables. In the following descriptions, we again assume a model producing disyllabic words, and refer to two frame node output ports, though, of course, frame nodes encoding the metrical structure of words with more syllables are also possible, where further ports would be required.

Asynchronous model. The first option to relax the equal duration constraint is to allow the durations of the periods of the activation pattern associated with each output port to differ. Thus, under this variant, the two ports are potentially out of sync relative to each other after word onset, because one parameter controls the durations of the activation periods output by the first port, and the other parameter controls the durations of the output of the second port. An example of a possible set of frame output patterns produced by this variant is depicted in the upper cell in Figure 3. This variant requires two parameters: dur_0 and dur_1. These control the duration in ticks of all phases of the output of Port 1 and Port 2, respectively.

Synchronous model. Alternatively, synchronization between the activation patterns could be maintained, such that when Port 1 is outputting full activation, Port 2 is outputting partial activation, but allowing the durations of each pair of steps to differ. This means that both ports always switch activation level at the same moment, but the amount of time that elapses between these switching events may vary. An example of a possible set of frame output patterns produced by this variant is shown in the lower cell in Figure 3. This variant has four duration parameters: dur_0, dur_1, dur_2, and dur_3, defining the duration of four phases that occur simultaneously in both output patterns—that is, the parameters all have influence on the activation patterns emitted from both ports. The parameter dur_0 defines the duration of the first phase, where Port 1 outputs partial activation and Port 2 outputs baseline activation. The duration of the second phase, where Port 1 outputs full activation and Port 2 outputs partial activation is specified by dur_1. The duration of the third phase, where Port 1 outputs baseline activation and Port 2 outputs full activation is defined by dur_2. The duration of the final phase, where both ports output baseline activation, is defined by dur_3.

Control model. We also constructed a control model variant that retains the timing structure described by Dell et al. (1997). The model variant performed poorly relative to the asynchronous and synchronous model variants, as expected. Full details about the control model variant are available in the online supplemental materials.

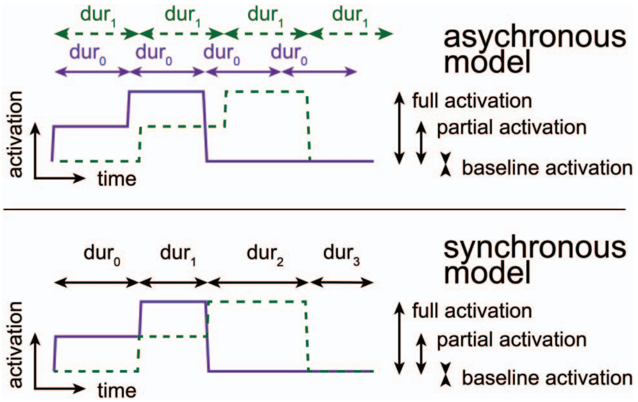

Figure 3. The activation patterns produced by the frame node for Port 1 (purple, solid) and Port 2 (green, dashed) in the asynchronous and synchronous model variants. The duration of each step in the activation patterns is controlled by various parameters, depending on the model variant (such as dur_0, see text for full details). See the online article for the color version of this figure.

Execution Component

To calculate the onsets and offsets of each syllable, we need to connect a model of the execution phase of speech production to the formulation phase. Our conception of the execution phase is straightforward; we assume that the duration of strong activation of a syllable output node is linearly related to the duration of articulation of that syllable (cf. Tourville & Guenther, 2011). To identify strong activation, we compare the activation of each syllable node over time to a syllable-specific threshold. Syllable-specific thresholds are used to account for variability in the magnitude of activation change in each syllable position. When the activation first exceeds this threshold, we consider syllable production to start, and when it decreases below the threshold again, we consider syllable production to stop. This procedure is fully specified the methods for Strand 1 (Evaluation of a Solution), and is functionally equivalent to assuming that execution faithfully reproduces the temporal dynamics of formulation, and that continuing activation from formulation is necessary during articulation.

Computational Implementation

The EPONA model is programmed in Python 3, using the NetworkX library (Hagberg, Schult, & Swart, 2008, Version 1.11), in which nodes and connections between them are defined and the spread of activation from node to node can be computed as a function of time. The optimization and learning of the model is also programmed in Python, using the Platypus library (Hadka, 2017, version as of April 2017). The code for the model and all analyses is available as part of the online supplemental materials and at https://osf.io/3mqgu/.

Speech Corpus

The model requires speech data to compare against. In this case, speech data were taken from the 'experiment 1' subset of the PiNCeR corpus gathered by Rodd, Bosker, ten Bosch, Ernestus, and Meyer (2019), which contains speech recordings and is annotated for word and syllable onset and offset times in ('CV.CVC and 'CCV.CVC) disyllabic Dutch words. The speech was elicited by means of cued picture naming, whereby 12 speakers named prefamiliarized line drawings presented in sets of eight on a "clock face" display. The words that were elicited are provided in the online supplemental materials. The picture to be named was indicated by a cueing dot, which moved clockwise from picture to picture, at slow (915 ms/word, 1.09 Hz), medium (646 ms/word, 1.56 Hz), and fast (456 ms/word, 2.19 Hz) rates. These speaking rates were selected on the basis of a pilot experiment where speakers were not cued, but instead encouraged to speed up or slow down as much as they could. These rates fall within the range of rates measured in the switchboard corpus of spontaneous speech (Greenberg, Carvey, Hitchcock, & Chang, 2003), but are all slower than the median rate in that corpus, and are slower than an estimate of mean rate for Dutch speakers of similar demographics (Quené, 2008). This is likely because the picture naming task, which included only middle-to-low frequent concrete nouns, was relatively hard compared to conversational speech, which includes many closed class words that are fast to plan.

The word onset and offset times were obtained by a multistep process. First, forced alignment using MAUS (Schiel, 2015) was applied to each trial (set of eight pictures). The resulting word boundaries were subsequently checked by a panel of experienced annotators, who evaluated whether the segmentation was accurate or not. Finally, the panel of annotators adjusted the boundaries of words that were marked as inaccurate in the previous step. Because the words were disyllabic, the onset of the first syllable and the onset of the word were simultaneous, and the offset of the second syllable and the offset of the word were simultaneous. To detect the onset of the second syllable, and the offset of the first syllable, a metric was employed to quantify the stability of the acoustic signal. Heightened acoustic instability was equated with temporal overlap between the gestural score encoding the first syllable and the gestural score encoding the second syllable. For further details about this metric, see Rodd, Bosker, ten Bosch, and Ernestus (2019).

The corpus contains 4,023, 3,575, and 2,627 word tokens for the slow, medium, and fast rate conditions, respectively. The size of the corpus sections differ primarily due to more frequent speaker error and less successful forced alignment in the faster conditions. However, within each speaking rate section, the remaining tokens were evenly distributed across the target words, and the proportion of 'CV.CVC versus 'CCV.CVC words was comparable between the corpus sections (29.7%, 29.9%, 30.6% 'CV.CVC words for fast, medium, and slow rates, respectively).[1]

The distributions of the first and second syllables and the overlap between them are shown for each cueing rate condition in Figure 4.

Training and Testing the Computational Model (Strand 1)

Strand 1 concerns the construction of a family of computationally implemented connectionist models of the formulation phase of speech planning, optimization of the model variants to mimic temporal properties of natural speech production, and evaluation of the performance of the model variants.

[1] Note that statistical testing to confirm whether or not the corpus sections differed is not appropriate, because the sets of words here are closed populations, rather than samples from some larger population (Sassenhagen & Alday, 2016).

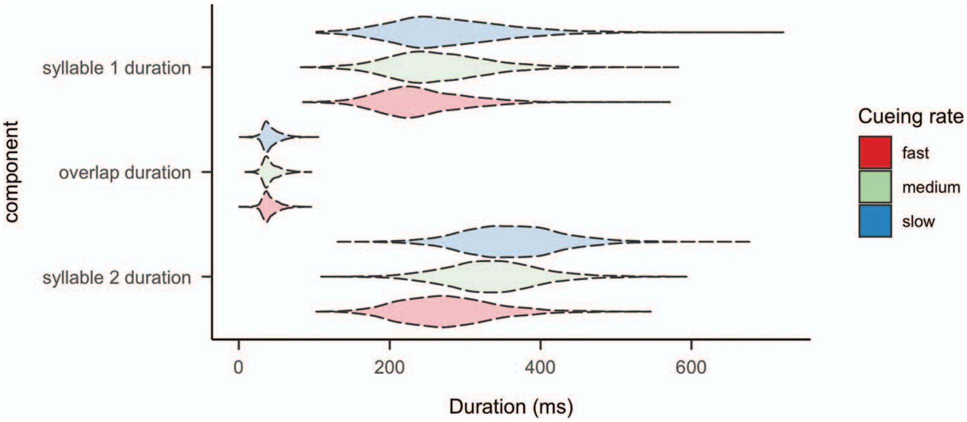

Figure 4. The distributions (violins) of the durations measured in the PiNCeR corpus, separated by rate condition. These form the three “fingerprint” distributions that the model seeks to mimic. See the online article for the color version of this figure.

Methods: Evaluating the Performance of Model Variants

Our aim in Strand 2 was to apply simulation of the EPONA to reveal how the cognitive system underpinning speech production can be modulated to achieve speech at different speaking rates. To achieve this, we require the model to simulate the performance of human speakers using different rates in Strand 1. However, it is not straightforward to evaluate how well a model simulates human speech production.

We consider the set of distributions of first and second syllable duration and overlap duration in each rate condition of the PiNCeR corpus as a “fingerprint” of the speech production system operating at that speaking rate (see Figure 4). Together, the fingerprint distributions capture more about the regimes of the speech production system than only the average durations of the syllables and the overlap between them would do, because the variation present in the durations is a product of variability inherent to the production system operating in a given regime. Because we are not concerned with individual differences between participants, but, instead with characterizing the regimes of the speech production system more broadly, we collapse across the 12 speakers when constructing the fingerprint distributions. The distributions of the data in the corpus shown in the violins in Figure 4 are therefore identical to the fingerprint distributions used to fit the models. Model optimization is then conducted independently for each speaking rate.

Optimization procedure. The Platypus (Hadka, 2017) implementation of the NSGAIII (Deb & Jain, 2014) algorithm was used to find the best parameter values in each speaking rate for each model architecture. The fitting procedure is depicted in Figure 5. The optimizer must find a set of parameter values that produce a prediction that is a good fit for all three fingerprint distributions simultaneously. In line with the optimization literature, we will term such a set of parameter values a *solution*. Because the model produces a distribution for each of the three fingerprint distributions, we obtain three estimates of fit quality for each solution tested: one for each distribution. In the optimization literature, such a quality estimate that is to be maximized or minimized is termed an *objective*. We obtain independent estimates of fit quality, in the form of the Kullback-Leibler (*KL*) divergence for each objective.

The *KL* divergence is a commonly used measure of the dissimilarity of two distributions, where a lower *KL* divergence indicates more similar distributions. By definition, its magnitude is dependent on the variability of the observed distribution. In our case, the variability of the observed duration distributions differs substantially between the three objectives. This means that the scales of the *KL* divergences calculated for each of the three objectives are not directly arithmetically comparable. We have no theoretical reason to prefer that the model concentrate on learning to fit one of the objectives ahead of the others, but simply summing (or averaging) the *KL* divergences would place undue weight on one of the objectives. We must therefore consider all three objectives together. Such an optimization problem with multiple independent estimates of fit quality (or objectives) that cannot be straightforwardly collapsed is known as a multiobjective problem. Typically, there is no single solution that is optimal for all objectives: Solutions that work well for one objective may be poor for another. Instead, the optimization algorithm aims to identify the solutions that are *Pareto efficient*, that is, the fit that they achieve for one objective cannot be improved upon without worsening the fit for one of the other objectives. This set of Pareto efficient solutions is termed the *Pareto front*.

Alongside the complication of multiple objectives, our models also have multiple free parameters to be optimized (between 11 and 14 depending on model variant; a full listing of parameters is available in the online supplemental materials), and are computationally expensive (time consuming) to evaluate because we simulate activation spreading through the network for each and every solution, and require multiple repetitions to simulate the fingerprint distributions. A complex error landscape with more than a handful of free parameters can prove difficult to search effectively; a classical method such as grid search, where evenly spaced points in the parameter space are sampled, requires prohibitively many model evaluations to get good coverage, and still runs the risk of missing good solutions between the sampled points. We suspected

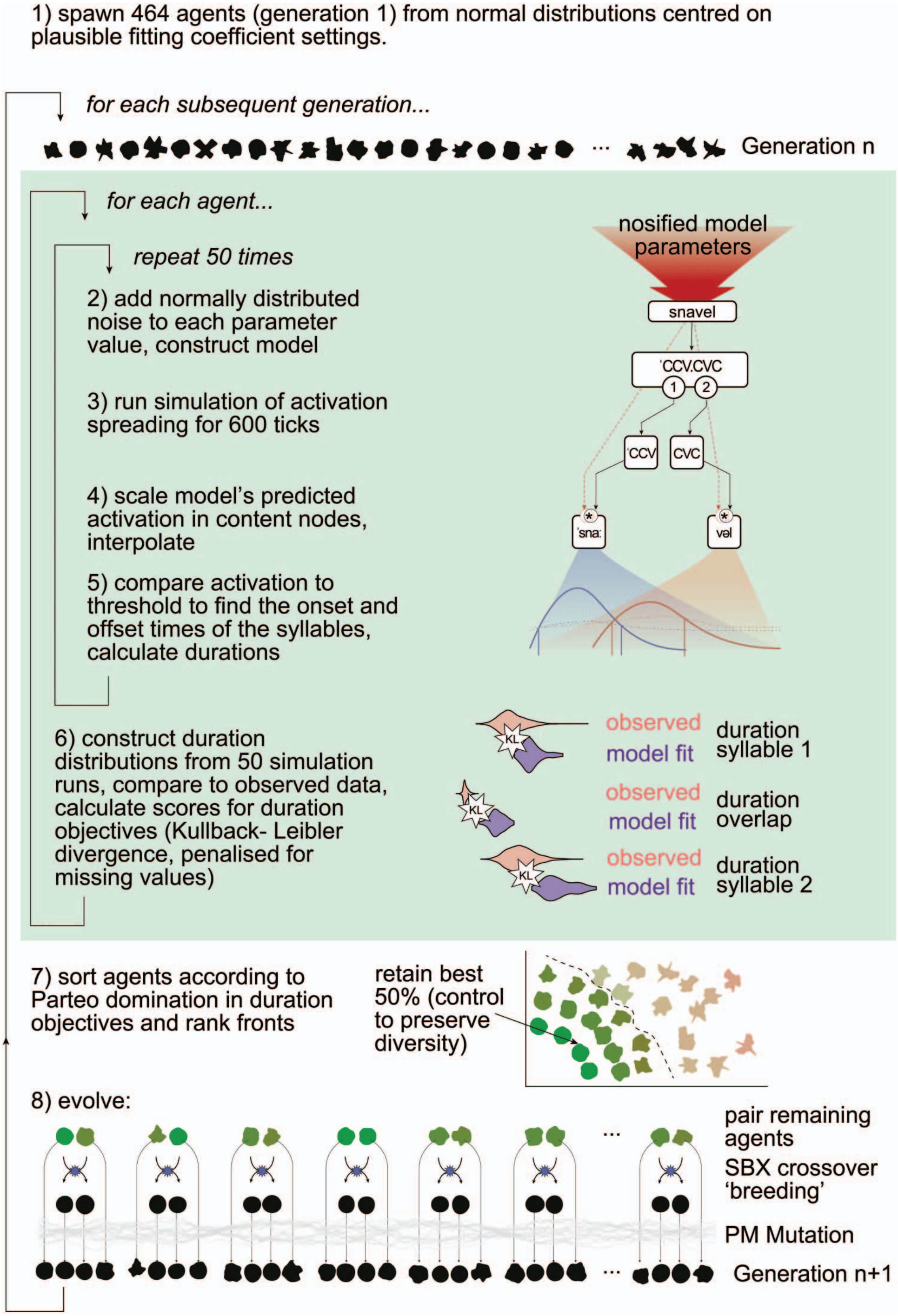
1) spawn 464 agents (generation 1) from normal distributions centred on plausible fitting coefficient settings.
for each subsequent generation...
...
Generation n
for each agent...
repeat 50 times
2) add normally distributed noise to each parameter value, construct model
3) run simulation of activation spreading for 600 ticks
4) scale model's predicted activation in content nodes, interpolate
5) compare activation to threshold to find the onset and offset times of the syllables, calculate durations
nosified model parameters
snavel
ˈCCV.CVC
1
2
ˈCCV
CVC
*
ˈsnaː
*
vəl
6) construct duration distributions from 50 simulation runs, compare to observed data, calculate scores for duration objectives (Kullback- Leibler divergence, penalised for missing values)
KL
observed
model fit
duration syllable 1
KL
observed
model fit
duration overlap
KL
observed
model fit
duration syllable 2
7) sort agents according to Parteo domination in duration objectives and rank fronts
retain best 50% (control to preserve diversity)
8) evolve:
...
pair remaining agents
SBX crossover 'breeding'
PM Mutation
...
Generation n+1

Figure 5. A diagrammatic representation of the fitting process. See the text for full details. See the online article for the color version of this figure.

that our parameter space might be quite complex, containing multiple clusters of good solutions in each rate condition. For these reasons, we selected NSGAIII. NSGAIII belongs to a class of optimization algorithms that accumulate knowledge about the search space over time (in multiple "generations" of learning "agents"). This means that the search can become gradually more focused on promising regions of the space. NSGAIII combines the ability to solve multiobjective problems with active preservation of diversity in the solutions it retains from generation to generation, making it suitable to search a space with many local minima. Other search methods such as particle swarm optimization have a tendency to converge early: that is, they are poor at exploring spaces where there are local minima (plausible solutions that are good, but not as good as the best solution in the space) at different positions (Kennedy, 2011; Peer, Bergh, & Engelbrecht, 2003).

In the remainder of this section, we will discuss the workings of the evolutionary algorithm in more detail, and then describe the procedure for evaluating the fit of the models, and the procedure employed to train the models.

Evolutionary algorithm. The NSGAIII (Deb & Jain, 2014) algorithm mimics evolution in biology. The evolutionary process begins by spawning a population of agents. An agent is a carrier of a "genome" (set of parameter μ and σ values, which define the central tendency and spread of distributions associated with each model parameter) that interacts with other agents to explore the parameter space. In each generation, the genome of the agent is varied somewhat by processes of mutation. Therefore each agent tests a different solution in each generation.

At the start of the optimization procedure, we spawn 464 agents, a population size recommended by the Platypus package based on the number of free parameters of the model variant with the most free parameters (Hadka, 2017). For the first generation, the parameter μ values for each agent (that agent's genome) are sampled from relatively broad normal distributions centered around values that we identified in pretesting as producing plausible activation sequences (Step 1 in Figure 5).

The model is then evaluated using the parameter μ and σ values associated with each agent for that generation, resulting in a fitness score for each fingerprint distribution for that solution. The simulation of the model and the procedure for evaluating a solution are described in the next subsection (Evaluation of a solution) and depicted as Steps 2 to 6 in Figure 5. The fitness scores are Kullback-Leibler divergences between the observed and predicted fingerprint distributions.

Once all agents in the generation are evaluated the Pareto optimal solutions are selected. Formally, a solution b can be said to dominate another solution a (denoted $a < b$) if it has a lower score on at least one objective while not having a higher score than a on any objective. The Pareto front is therefore the set of solutions that are not dominated by any other solution. The solutions of this first "Pareto front" are assigned a rank of 0. From the remaining unranked population, a new set of solutions that are Pareto optimal in the smaller population are identified, and assigned a rank of 1. This procedure is repeated to find subsequent fronts, with the agents in the third front being assigned a rank of 2, and so on, until all agents are ranked.

The agents are then entered into selection "tournaments," in which two agents are randomly drawn from the population, compared, and the agent with the lower rank is retained. The losing agent is discarded from the population and no longer contributes to future generations. Further tournaments are performed until all agents have competed once (Step 7 in Figure 5). This has the advantageous effect that all agents from the best rank will be retained, all agents from the worst rank will be excluded, and that agents from the ranks in between have a gradually decreasing chance of being retained. This, along with a further mechanism to preserve agents in underrepresented parts of the parameter space (Deb & Jain, 2014, p. 582), means that the retained agents represent, broadly, the best half of the initial population, but that, simultaneously, variability is maintained, which ensures that the optimization procedure searches the "bumpy" parameter space effectively.

Then, the evolution stage begins (Step 8). The remaining agents are randomly paired up and recombined to make offspring by the simulated binary crossover operator (Deb & Agrawal, 1995; Deb, Sindhya, & Okabe, 2007), which simulates the mixing of two genomes in sexual reproduction. For each pair of parents, for each value in the set of parameter μ or σ, a polynomial probability distribution is constructed around each parent value. Two sets of child values are then sampled from the mixture distribution (Deb & Agrawal, 1995). This results in child agents that combine traits from each parent agent. The parents and the children together form the population for the next generation of evaluation, competition, and recombination, after having been subjected to further random mutation by the polynomial mutation operator, where a perturbation is sampled for each parameter μ or σ value from a polynomial distribution centered at zero (Deb & Agrawal, 1995; Deb & Goyal, 1996). Because of this mutation step, specific solutions are usually not repeated in subsequent generations, and the overall fitness of a next generation may be worse than a previous generation, but in general the optimization procedure will result in improved scores over time. In our implementation, 5,000 generations were run.

Evaluation of a solution. The process of evaluating the set of parameter μ and σ values associated with an agent is illustrated in Figure 5 (in the green box), and described in detail below. The aim of this evaluation procedure is to assess how well each set of parameter μ and σ values mimics the observed fingerprint distributions. This requires us to construct a distribution of each of these variables.

To construct the predicted distributions, we run the model 50 times with each set of parameter μ and σ values. In each of the 50 repetitions, a small amount of noise is added to each parameter μ value, sampled from a normal distribution centered at 0, the standard deviation of which is defined by the parameter σ value. These noisified model parameters are used to construct an instance of the model variant to be tested, with node properties and connection weights defined by the model parameters (Step 2; see the online supplemental materials for a full listing of the model parameters).

The model keeps time internally using a unit that is 9 ms long, a "tick"; activations are recalculated once per tick. This value was arrived at by pretesting with models where the number of ms that each tick represents was learnt along with the other parameters. In the simulations reported here, the duration (in ms) of a tick was held constant across word productions. A unit of this order of magnitude is convenient because it allows sufficiently detailed sampling (e.g., the shortest segments are still represented by several ticks) but allows faster computation than a shorter tick length (cf. typical window shift of 10 ms in MFCC measurements, Young et al., 2006).

Each model is run for 600 ticks (that is, we calculated the activations in the network 600 times) which amounts to 5,400 ms, a duration long enough for the word to be produced and the activation of all nodes in the network to return to baseline, whatever the speech rate condition.

Activation of the plan node always occurs after four ticks, and persists for 28 ticks at a constant activation level determined by a model parameter. After 28 ticks, the activation in the plan node decays, at a decay rate determined by a model parameter. These values were also arrived at during pretesting, where these parameters were allowed to vary. Holding these parameters constant across conditions ensures that the differences between speech rates emerge in the nodes contained in our model, rather than resulting from higher level processes that we assume to be responsible for activating the plan nodes. The activation in the plan node spreads through the nodes of the network, finally reaching the content nodes (Step 3). The time courses of the activation in the content nodes of the model are extracted, and the resulting time courses are linearly interpolated every 0.1 ticks (Step 4), yielding time courses k_{t_s}, for time t and syllable s.

Next, we need to establish the times where we suppose that the activation in the content nodes is sufficient to result in production of the syllable. We do this by comparing the interpolated activation time course k_{t_s} against a separate threshold θ_s for each syllable s. The threshold for each syllable is calculated as the sum of a constant which is the same for all syllables, and a weighted exponential moving average of previous activations in the relevant content node. This means that the threshold gradually increases in response to activation in the content node, mimicking short term adaptation to the activation.

To calculate the threshold, we need to calculate the moving average activation. We calculate the average over a Gaussian kernel. First, a weighting factor α is calculated, to cause the moving average activation to operate over a span of nine ticks (90 observations with one observation every 0.1 ticks). The moving average activation m_{t_s} at a given time t for a given syllable s is then calculated recursively from the activation time series k_{t_s}:

$$m_{t_s} = \begin{cases} k_t, & t = 1; \\ \alpha k_{t_s} + (1 - \alpha)_{t-1}, & t > 1. \end{cases} \qquad \alpha = \frac{2}{90 + 1} = 0.022 \tag{1}$$

Then, the threshold θ_{t_s} is calculated as the sum of the offset u, which is a model parameter, "threshold_constant," and the moving average activation m_{t_s}, multiplied by a weighting ($c = 0.1$, for all conditions):

$$\theta_t = u + cm_t \tag{2}$$

The moment when the activation in the first syllable content node exceeds its threshold is taken as the onset word production, and the time when the activation falls below the threshold again is taken as the offset of the first syllable. The moment that the activation in the second syllable content node exceeds its threshold is taken as the onset of the second syllable, and the time when the activation falls below its threshold is taken as the offset of word production (Step 5). In some instances, the model may predict multiple periods or activation for a syllable, or no activation at all. In cases where there is not precisely one period of activation above the threshold for each of the two syllables, no onset or offset times are recorded for that repetition. This suggests that the set of parameters is not very robust, and is excessively sensitive to the subtle changes introduced by the noisification, and should be dispreferred by the optimization algorithm.

From the syllable-level onset and offset times, the three objectives can be calculated for each repetition: syllable 1 duration, between-syllable overlap, and syllable 2 duration. The durations from each of the 50 repetitions (n_{reps}) are collected and a predicted distribution is constructed (Step 6). To score the quality of the fit achieved by the values of the parameters, the observed fingerprint distributions p are compared with the predicted distributions q, for each fingerprint duration objective obj (Step 7). The predicted and observed distributions are first binned (bin width 8 ms, from −200 ms to 1,000 ms relative to simulation onset, 150 bins, n_{bins}), and a constant floor value ε of 1×10^{-13} is added to the count in each bin. The count in each bin is then divided by the sum of the counts in all the bins:

$$p_{obj_b} = \frac{\text{count predicted}_{obj_b} + \varepsilon}{\sum_{b=1}^{n_{bins}} \text{count predicted}_{obj_b} + \varepsilon} \qquad q_{obj_b} = \frac{\text{count observed}_{obj_b} + \varepsilon}{\sum_{b=1}^{n_{bins}} \text{count observed}_{obj_b} + \varepsilon} \tag{3}$$

Then, the Kullback-Leibler divergence is calculated:

$$KL(p_{obj}, q_{obj}) = \sum_{i=1}^{n_{bins}} p_{obj_i} \times \log_2\left(\frac{p_{obj_i}}{q_{obj_i}}\right) \tag{4}$$

where p is the observed distribution and q is the predicted distribution. $KL(p_{obj}, q_{obj})$ is taken as the score for the objective obj.

In cases where not all of the 50 simulation repetitions resulted in a duration (because the onsets and offsets of the syllables stray outside the period of the binning, because the activation time series never crosses the threshold, or because the activation times series crosses the threshold multiple times), the score was penalized by multiplying the KL by 50 (the number of repetitions, n_{reps}) divided by the number of values present. This penalization is intended to favor solutions that are more stable; that is, all 50 repetitions predicted exactly one period of activation for each syllable:

$$\text{missing}(p, q) = n_{reps} - \sum_{b=1}^{n_{bins}} count(p, q)_b \tag{5}$$

$$score_{obj} = \begin{cases} KL(p_{obj}, q_{obj}) \times \dfrac{n_{reps}}{n_{reps} - \text{missing}(p, q)}, & \text{missing} < n_{reps}; \\ KL(p_{obj}, q_{obj}) \times n_{reps} * 1.2, & otherwise. \end{cases} \tag{6}$$

Learning procedure. To test the models, two phases of optimization were conducted for each model variant for each rate condition. During the first 100 generations of the optimization procedure, some of the parameters are clamped; that is, the algorithm does not adjust them. This phase can be thought of as a rough initial search of a dimensionally reduced subset of the parameter space. After this phase, the clamping of these parameters is released, and all the parameters are fine tuned to optimize the model's output. A full listing of the parameters, indicating which are clamped during the first 100 generations, is available in the online supplemental materials. The optimization procedure is run for another 900 generations. During the first 1,000 generations, the σ associated with each parameter is linearly related to the parameter μ value ($\sigma = 0.08 \times \mu$), following the observation of a linear relationship between the center and the spread of the distribution in, for instance, response times (Luce, 1986; Wagenmakers & Brown, 2007).

After the 1,000th generation, clamping is applied to most of the parameter μ values, such that they no longer undergo changes during the evolution and mutation phases of the NSGAIII algorithm (see the table in the online supplemental materials for full details), while the parameter σ values are released, and therefore learnt independently. Starting with the 1,500th generation, this is reversed, and the σ values are clamped and μ values are learnt. Starting with the 2,000th generation, σ values are again released from clamping, and μ values are clamped. From the 2,500th generation, no clamping is applied. The learning procedure is stopped after the 5,000th generation. (This arrangement is depicted graphically in the shading in Figures 6 and 7.)

This multiphase approach is an attempt to speed up the overall search for a well performing parameter set, by allowing quick rejection of unpromising areas of the parameter space during the first 100 generations, and successively finer-grained searching in the subsequent phases. The long total run, of 5,000 generations, ensured that the optimization process was sufficiently converged to make valid model comparisons.

Explicit Test of the Advantage of Nonlinearity

It is also possible to vary the fitting procedure to more directly assess the hypothesis of the presence of gaits manifested as qualitatively different regimes in the parameter space. This "linear constraint" model is functionally identical to the asynchronous model variant, but is optimized in a different manner, to force the parameter values found during the optimization routine to be linearly related. Instead of conducting an independent optimization run for each rate condition, the parameters of the linear constraint model associated with all three speaking rates are optimized together via a metamodel. This metamodel has parameters for the slope of a line for each of the model parameters, as well as an intercept parameter for each speaking rate. From these slopes and intercepts, parameter values for each speaking rate are derived, and passed to instantiations of the asynchronous model variant for each speaking rate. The *KL* scores for each metric are gathered from the submodels, and together form the nine objectives (syllable 1 duration, syllable 2 duration, and overlap duration for each of the three speaking rates) of the multiobjective optimization routine. For clarity and conciseness, the results obtained from this additional model variant are reported along with those of the other model variants in the next section, where the model variant is referred to as the "asynchronous model variant with linearity constraint."

Results: Model Performance

Conventionally, statistical comparison of models for the purpose of model selection takes into account the number of parameters (degrees of freedom) that each model has; assigning models a "handicap" per extra degree of freedom to identify the model that strikes the best balance between quality of fit and parsimony (Akaike, 1974). In a framework where a model predicts variance, it is fairly clear how one would go about doing this. In our case, however, the models predict the three fingerprint distributions, which we evaluate on the basis of the Kullback-Leibler divergence between the model and the observed fingerprint distributions, rather than predicting values for each observation, from which likelihood-based metrics might be calculated. This makes it difficult to select a plausible handicap with which to penalize the model performance without adding further simulations.

A typical approach to assess the performance of different variants of a model is to directly compare their ability with fit the data after learning, by seeing how well the target function is satisfied by each trained variant. In our case, this is not possible because of the

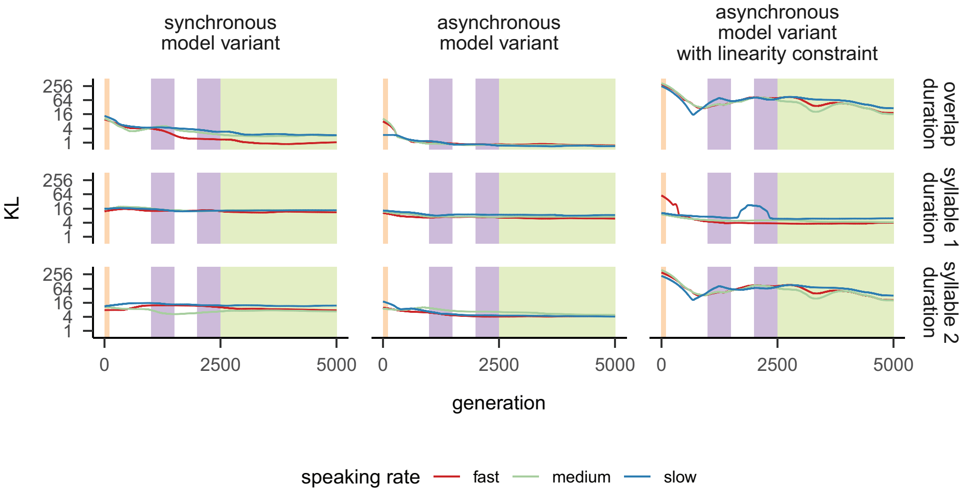

Figure 6. Loess-fits of the Kullback-Leibler scores (*y*-axis, log-transformed scale, lower values indicate better performance) of the solutions in the Pareto front in each generation (*x*-axis), for the three rate conditions (line colors), the three objective functions (rows), and three model variants (columns). The shading indicates the optimization phases of the model, orange is the phase where only the μ component of a subset of the parameters was adjusted by the optimizer, white indicates that the μ component of all parameters was adjusted by the optimizer, purple indicates that the σ component of all parameters was adjusted by the optimizer, and green indicates that both μ and σ components of all parameters were adjusted. See the online article for the color version of this figure.

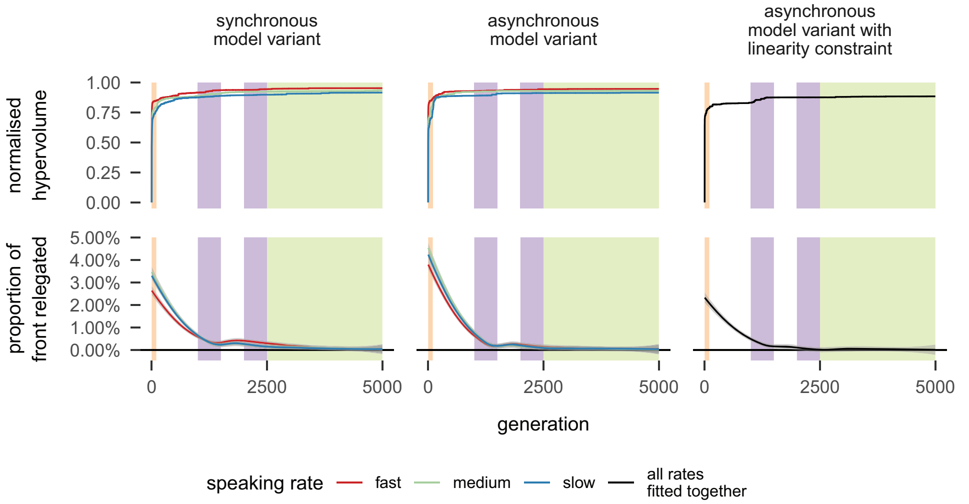

Figure 7. Upper panels: The normalized hypervolume indicator (*y*-axis) during the 5,000 generations of the optimization run (*x*-axis), for the three model variants (columns). Stabilization of the normalized hypervolume indicator at a value close to 1.0 indicates successful convergence. For the synchronous and asynchronous model variants, colored lines indicate the speech rate condition being optimized. Lower panels: the proportion of former front members relegated from the front in each generation. See the caption of Figure 6 for the meaning of the shading. See the online article for the color version of this figure.

multiobjective nature of the problem. Recall that the model optimization process results in a Kullback-Leibler score for each of the target distributions for each solution and that these scores are not mathematically comparable across the three objectives without unduly favoring one objective above another. We therefore needed to take a different approach to ascertain how well the different model variants learned, and how well they ultimately performed after training, that would not arithmetically collapse the Kullback-Leibler scores. To assess learning over time, we adopt a metric in terms of Pareto dominance. To assess final performance, we adopt a regression approach.

Learning trajectories. To characterize the learning trajectory of each run, we identified the Pareto front in each generation cumulatively. This means that, for each generation, we looked for solutions in that generation and all generations before it that were Pareto optimal. We used loess-fitting (Cleveland & Devlin, 1988) to identify the trend in the score for each objective function in each rate condition. These loess-fits are shown in Figure 6, where we can observe, very broadly speaking, that for all three model variants, the most progress is made in finding solutions that improve the fit in the overlap duration objective. Much more restrained progress is made on improving the fit of the syllable duration objectives. The asynchronous model appears to perform moderately better than the synchronous variant on the syllable duration objectives, while the variant with the linearity constraint never achieves scores as good as the other two variants, with the notable exception of the syllable 1 duration objective, which performs comparably to or slightly better than the other model variants.

Convergence. If the model is learning, the quality of the Pareto front will improve with each generation. Conventionally, convergence in the optimization multiobjective problems is assessed with the hypervolume indicator (Zitzler, Brockhoff, & Thiele, 2007), which calculates the volume of the dominated space between a reference point and the Pareto front. The hypervolume indicator for our optimization runs, normalized to have a value between 0.0 and 1.0, is presented in the upper panels of Figure 7. The value of the normalized indicator increases as the volume of the dominated space increases. Convergence is evidenced by stabilization of the indicator at a value close to 1.0.

Although simple to interpret and widely applied, the hypervolume indicator has the disadvantage of arithmetically combining the values of the objective functions into a single fit quality metric. This is undesirable for our *KL* objective functions. We therefore calculated a second indicator of model convergence, which assesses the change in the composition of the Pareto front after each generation.

When the model finds a new solution that is nondominated, this solution joins the Pareto front. Sometimes, this solution falls between two others, improving the coverage of the Pareto front, but not improving the fitness of the Pareto front in general. Other times, the solution dominates a solution or several solutions that were in the Pareto front in the previous generation. These dominated solutions are “relegated” from the Pareto front. Because we are primarily interested in finding optimal parameters to fit the observed data, and only secondarily interested in increasing the size of the Pareto front, we want a metric that is sensitive to the second type of new solution. Therefore, rather than counting new solutions, we count the number of solutions that are relegated from the Pareto front (cf. Martí, Garcia, Berlanga, & Molina, 2009). When the optimizer has converged, no relegation events will be observed. The lower panels of Figure 7 show loess fits of the proportion of former Pareto front members that are relegated in each generation.

Both the hypervolume indicator and the relegation count metrics indicate stability after around 3,000 generations, leading us to conclude that the optimizers are sufficiently converged by the end of the 5,000 generations tested.

Statistically testing model variant performance. In order to evaluate the performance of the different model variants, we need to identify and statistically test differences in the *KL* scores achieved by the Pareto front solutions of each of the model variants. Simultaneously, we need to disregard variation in the *KL* scores as a function of objective, because *KL* scores for the various objectives are not directly arithmetically comparable because of differences in the observed distributions, as previously discussed. The same holds for comparing models fitting different rate conditions, between which there are also differences in the variability of the observed distributions.

Instead of averaging scores across objectives, linear regression with categorical predictors for model variant, rate condition and objective can be used to isolate the effect on the *KL* score attributable to model variant, independent of rate condition and objective. This leads to a regression model with the following structure (Wilkinson-Rogers notation, Wilkinson & Rogers, 1973):

$$KL \sim \text{model variant} * \text{rate condition} * \text{objective} \qquad (7)$$

This is a model predicting *KL* with categorical predictors for *model variant*, *rate condition*, and *objective*, and all interactions between the levels of those categorical predictors.

The *KL* scores were bootstrap resampled to introduce variation required to perform regression modeling. The bootstrapped distributions of the *KL* scores are shown in the first three panels of Figure 8. We took 2,000 samples with replacement of sets of syllable 1 duration, syllable 2 duration, and overlap duration values from the observed dataset. For each of these samples, we calculated the *KL*s between the resampled observed distributions and the model's predicted distributions. The resulting bootstrapped *KL*s were then log transformed and z-normalized. The log transformation was necessary to de-skew the *KL*s, which obey a log distribution.

The regression model fitted the data quite well, achieving an adjusted R^2 value of 0.76. The fits of the regression model for the main effect of model variant are shown in the fourth panel of Figure 8, as black dots. The full table of model coefficients is provided in the online supplemental materials.

Relative to the asynchronous model variant, the synchronous model variant performed significantly worse (β = 0.55, SE = 0.0083, t = 66^{***}, d = 0.52).

As discussed earlier in this section, it is not possible to draw meaningful conclusions from the significance of the main effects of rate condition or objective; these were included to enable us to

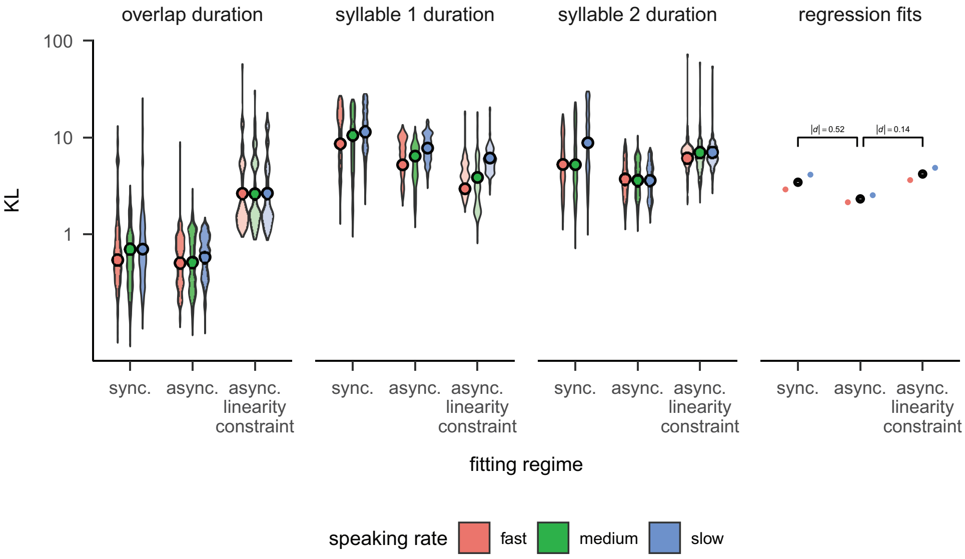

Figure 8. First three panels: the bootstrapped distributions (violins) of the KL scores (*y*-axis, smaller is better, log scale) achieved by the 0 ranked agents (the Pareto front) for each model variant (*x*-axis, *sync.*: synchronous model variant, *async.*: asynchronous model variant and *async. linearity constraint*: asynchronous model variant with linearity constraint, see the text for full details) in each speaking rate condition (fill colors), in each objective (panels). The colored dots indicate the model fits for the three-way interaction term in the regression model. Fourth panel: the fits of the model variant term from the regression model (main effect shown as black dots, fits of rate condition:model variant interaction in smaller colored dots). Ninety-five percent confidence intervals are omitted because they are too small to be visible. Significant differences in the main effect are indicated. The main effect of model variant is plain to see; the asynchronous model variant performs significantly better (achieves lower *KL* scores) than the synchronous model variant. The asynchronous model variant without the linearity constraint outperforms the asynchronous model variant with the linearity constraint. See the online article for the color version of this figure.

use the regression model to avoid arithmetically comparing *KL* scores calculated with different observed distributions and therefore different scales.

Are predicted fingerprint durations plausible? It is also informative to assess the performance of the model variants qualitatively, by directly examining their success in emulating the target distributions. In Figure 9, we show the distributions resulting from combining the duration distributions predicted by each member of the Pareto front of each run as solid violins. These are compared against the target distributions measured from the corpus (translucent violins with dashed edges).

For all three model variants, relatively good fits are achieved to the syllable 2 duration distribution, with the asynchronous model variant arguably mimicking the precise shape of the distribution somewhat better than the synchronous model variant and the asynchronous model variant with linearity constraint. In fitting the syllable 1 duration distribution, the synchronous model variant produces a bimodal distribution, rather than the unimodal distribution in the observed data, and also fails to fit the central tendency well. The asynchronous variant performs better, although the distributions it predicts are slightly too narrow. The asynchronous variant with linearity constraint predicts syllable 1 duration distribution very well. In fitting the overlap duration distribution, the asynchronous model variant performs best, fitting the central tendency well but overestimating the spread of the distribution somewhat. The asynchronous model variant with the linearity constraint predicts a slightly wider unimodal distribution. The synchronous model variant again predicts a bimodal distribution where one mode matches the density peak of the observed distribution.

It should be noted that the vast majority of simulation papers in this domain report only central tendencies. The distributional fits that we achieve seem acceptably good in (qualitative) comparison with the few psychological modeling studies that we found that did fit distributions (Wiecki & Frank, 2013, Figure 4; Engbert, Nuthmann, Richter, & Kliegl, 2005, Figure 10).

Summary of Strand 1

In Strand 1 of this study, we introduced EPONA, a new model inspired by the DBS model, that was successful in predicting the temporal structure of disyllabic word production. EPONA provides the first computationally explicit connectionist account of speakers' ability to modulate the speech production system to achieve different speaking rate.

The methods that we used to train and evaluate the variants of the model were also novel. We adopted a framework whereby the model predicts distributions of three objectives, which were measured from the PiNCeR corpus of elicited speech (Rodd et al., 2019): the duration of the first syllable, the duration of the inter-syllable overlap, and the duration of the second syllable. We assumed that the central tendency and the variability of these distributions together reflect the characteristics of the underlying cognitive system. This means that during the training process, the models learned to resemble the underlying cognitive system.

Training proceeded using an evolutionary algorithm that optimized the parameter values so as to minimize the Kullback-Leibler divergence scores associated with each objective distribution. The success of the evolutionary algorithm in learning parameter values that fitted the objective distributions for each model variant is an index of how well suited that model variant is as a model of the formulation phase of speech production.

Alongside the asynchronous and synchronous model variants, we introduced a third model variant that was fitted using a different optimization regime. This allowed us to directly test the prediction of a single gait system, where all three speaking rates are linearly related in parameter space. This model is discussed further in Strand 2.

The asynchronous model variant without the linearity constraint performed best on the quantitative criteria we set and offered the most plausible predicted fingerprint durations. We therefore perform further analyses for Strand 2 only on the asynchronous model variant.

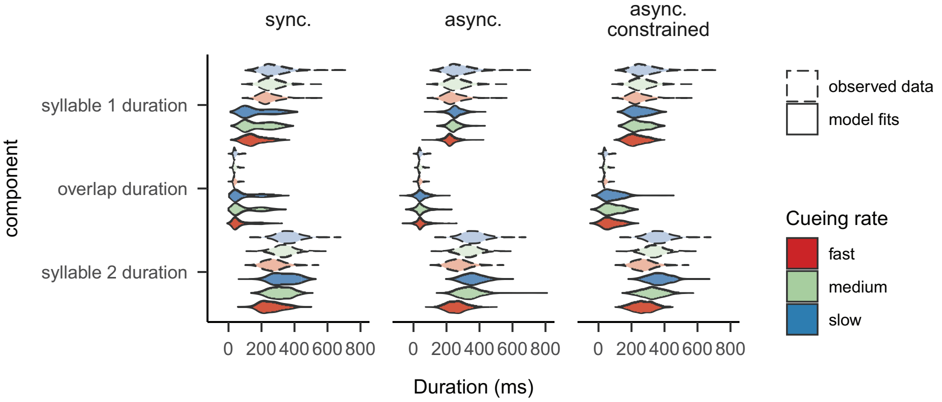

Figure 9. The duration (*x*-axis) distributions (filled violins) predicted by three models variants (facets) at the three rate conditions (colors) for each of the three target distributions (*y*-axis), compared against the observed distributions (translucent violins with dashed edges). See the online article for the color version of this figure.

How Do Regimes Relate to Each Other? (Strand 2)

To explore how executive control might be exerted on the EPONA model to achieve different speech rates, and thereby assess whether different rates are achieved by shifting between multiple qualitatively different "gaits" of speech production, we need to compare the best parameter values identified by the optimizer for each speaking rate condition. We can think of the solutions as positions in a multidimensional space where each parameter of the model is mapped to one dimension. In such a space, the Euclidean distance between a pair of locations in parameter space represents the difference between solutions.

Note that we have assumed that only one regime exists for each speaking rate, while of course several distinct configurations may have emerged to account for the temporal structure of speech at a given rate. We tested for this possibility by performing k-means clustering on the parameter values associated with each speaking rate. The clustering did not support multiple regimes in any of the rates; see the online supplemental materials for full details.

How Are Regimes Arranged Relative to Each Other?

Method. Having identified the best solutions for each rate, we consider how the regimes adopted for each rate relate to the regimes adopted for the other rates. To do this, we perform principal component analysis (PCA), which involves projecting the 12 parameters on which the speaking rate regimes vary onto principal components (PCs). The procedure loads as much variance as possible onto each component in turn, while ensuring that each component is orthogonal to the preceding PCs. A full listing of the parameters is provided in the online supplemental materials. PC1 (the first PC) accounted for 30.0% of the variance, PC2 accounted for 11.6% of the variance, PC3 for 8.6% and PC4 for 5.3%. The loadings of the parameters onto the PCs are listed in the online supplemental materials.

Results. Figure 10 shows the spread of solutions across the rate conditions in the first and second principal components. Note that because this is a projection of multiple dimensions into two, much variation is not visible, and points that appear adjacent on the PC1-PC2 plane depicted may be quite distant on other dimensions. For this reason, it is not certain that medium and slow are closer together than medium and fast, or slow and fast, although it appears so on the PC1-PC2 plane. The optima associated with the three rates (fast in red, medium in green, and slow in blue tones) occupy broadly different areas of the PC1-PC2 plane. On this plane, the clusters of solutions of the three conditions are well separated, and the spread of the solutions in the three conditions is broadly comparable.

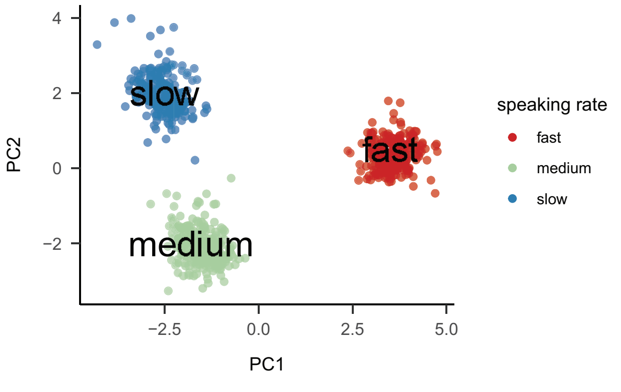

Figure 10. The Pareto optimal solutions identified for the fast (red), medium (green), and slow (blue) rate conditions, plotted for PC1 (*x*-axis) and PC2 (*y*-axis). See the online article for the color version of this figure.

The spatial organization of the rate conditions on the PC1-PC2 plane is clearly not axial in nature, ruling out the single gait account. This is in line with the observation in Strand 1 that an asynchronous model variant constrained to only consider linear arrangements of the rates in parameter space performed worse than the asynchronous variant without this constraint. Instead, the gaits are arranged as a triangle, supporting a multiple gait interpretation. Decelerating from the medium speaking rate to the slow speaking rate involves increasing PC2 while slightly decreasing PC1. Accelerating from the medium speaking rate to the fast rate involves increasing both PC1 and PC2.

Which Regimes Belong to Which Gaits?

Extrapolating fingerprint durations between rate centers. The previous finding suggests that there is more than a single gait adopted by speakers to control their speaking rate. The parameter optimization analysis cannot, however, allow us to assess which, if any, of the three regimes belong to the same "gait." To assess that, we conducted a further exploratory analysis.

We calculated the mean position of each speaking rate regime in parameter space. These means form the "reference" points. Between each pair of reference points, we interpolated five equally spaced points along a straight line (axis) through parameter space. Additionally, we extrapolated two extra points on each of these axes beyond the reference points. We therefore have axes from fast to slow, from fast to medium, and from medium to slow, that intersect at the reference points. The arrangement is illustrated the upper panel of Figure 11.

We then took the parameter values associated with the location of each point, and constructed and ran new instances of the asynchronous model with these parameter settings, to predict the distributions of the three "fingerprint" durations. Just as in the optimization procedure, the parameters were noisified, and 50 runs were conducted (see Figure 5 and accompanying text for more details). These durations, along with the word duration are indicated in the raincloud plots in panel C of Figure 11, and normalized in panels D and E.

Previously, we identified five possible mappings of the speaking rate regimes onto one to three gaits (single gait, three gaits, slow is distinct while fast and medium are mapped to the same gait, fast is distinct, medium is distinct). These possible mappings are depicted diagramatically in panel B of Figure 11. We directly modeled and compared the plausibility of these five hypothetical mappings. If a pair of speaking rate regimes belong to the same gait, we would expect the fingerprint distributions of the interpolated points between them to follow a linear trend, and that all the interpolated points would result in plausible fingerprint distributions. If, however, the regimes belong to different gaits, we would expect to see a nonlinearity at some point along the axis, indicating a shift from areas of parameter space associated with one gait to areas of parameter space associated

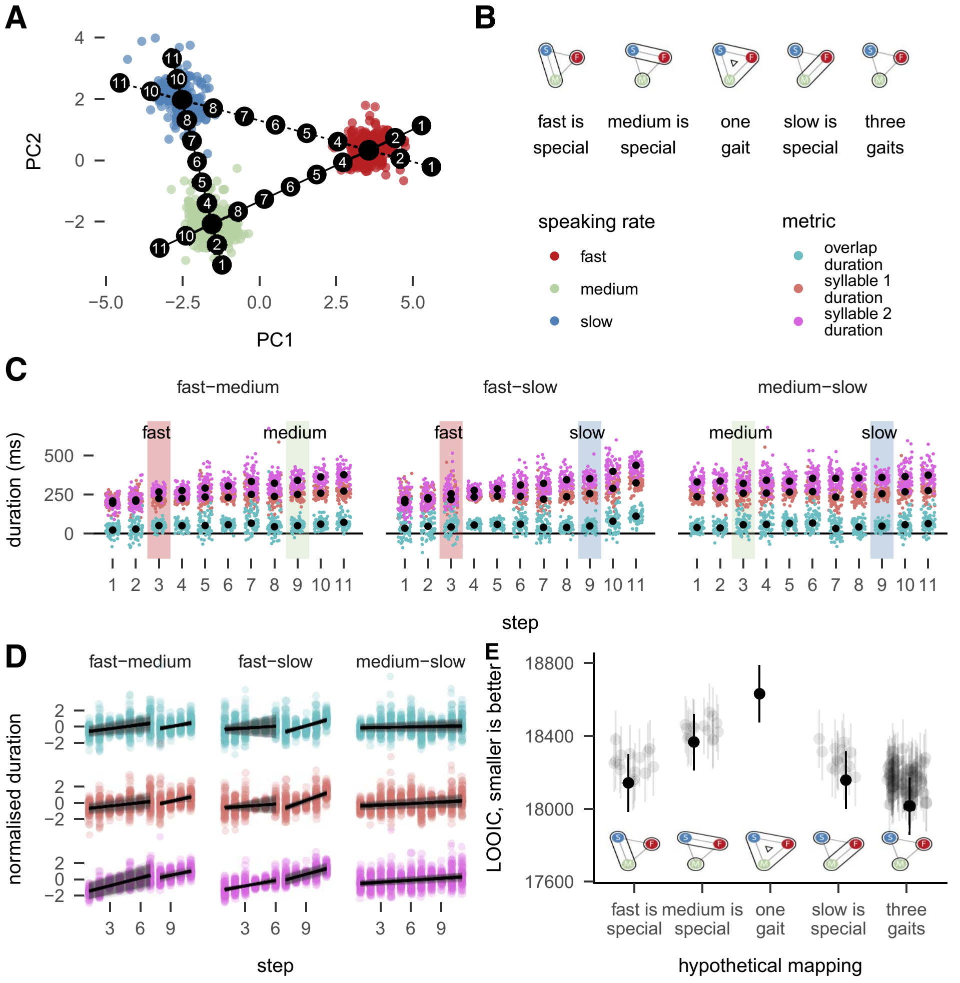

Figure 11. Panel A: The extrapolated axes projected onto the PC1 (*x*-axis)–PC2 (*y*-axis) plane through parameter space. Each black point indicates a location at which the model was run and fingerprint distributions were calculated. Behind, the optimal solutions identified by the optimization procedure are shown (see Figure 10 for details). Panel B: The hypothetical mappings of rates to gaits, represented diagramatically, enclosing lines indicate rates achieved by the same gait. Panel C: The distributions of the durations (*y*-axis), of the overlap, syllable 1 and syllable 2 (colors), shown as rainclouds at each step (*x*-axis) of the three axes (panels). Black points indicate the median values. Shading indicates the reference points where the axes intersect. Panel D: Example fit of the Bayesian linear switchpoint models for the “fast is special” mapping for the three axes (columns) and fingerprint component distributions (rows, colors). Panel E: Point estimates and standard errors of the quality of fit of the Bayesian linear switchpoint regression models for each mapping, quantified by an information criterion calculated by leave-one-out cross-validation. For each mapping, the models other than the best performing are plotted more lightly. See the online article for the color version of this figure.

with the other gait, possibly with an area of “unproductive” parameter space in between where nonplausible fingerprint distributions are predicted.

We tested the presence of linearity in the axes through statistical modeling of the simulated durations depicted in the lowers panel of Figure 11. We conducted both Bayesian (MCMC sampling) and non-Bayesian analyses using linear regression models and generalized additive models (GAMs, Wood, 2017). Both types of model were multivariate, in that they fitted the simulated durations of the three axes simultaneously in a single model. The results of the two approaches were comparable. For brevity, only the Bayesian analysis is reported here. The GAM analysis is reported in the online supplemental materials.

Bayesian linear switchpoint regression. For each axis of the extrapolated fingerprint duration data, we regressed the normalized fingerprint durations by the number of the step along the axis. The Bayesian models allow us to identify the locations in parameter space of the switchpoints along the axes, and additionally exploited variation in the distribution along the length of the axes.

Axes could be modeled with either a “uniform” linear fit, or a “switching” fit that permitted nonlinearity. The uniform fit predicted normalized duration (both μ and σ) by the step number, with distinct slope and intercept parameters for each Component × Axis combination for μ and σ. The switching fit split the axis into two halves at a fixed switchpoint, and fitted a separate regression with separate parameters for each half. For each axis, different fixed switchpoints were tested, namely between Steps 4 and 5; between Steps 5 and 6; between Steps 6 and 7; or between Steps 7 and 8. This means that different numbers of models were required for each mapping, ranging from 1 model for the "no gaits" mapping, to 64 models for the three "distinct gaits" mapping (4^3). A student *t* distribution was used as the likelihood. This has heavier tails than a normal distribution, meaning that it is a form of robust regression and can better accommodate heteroskedasticity. For all slope and intercept parameters, mild $N(0, 1)$ priors were applied, which makes the assumption that most effects are smaller than Cohen's $d = 1$ and nearly all effects are smaller than Cohen's $d = 2$. The fit resulting from the model fitting the "fast is special" mapping is depicted in panel E of Figure 11, by way of example.

For each model, eight chains of 8,000 samples (of which 4,000 warm up) were sampled by NUTS in RStan (Stan Development Team, 2018, Version 2.18.2). No convergence issues, assessed by the Gelman-Rubin diagnostic $\hat{R}$, effective number of samples, and visual inspection of traceplots were noted for any of the models. Full details of the Bayesian linear switchpoint analysis are available in the online supplemental materials.

Results. Panel E of Figure 11 presents the model comparison results of both the Bayesian linear switchpoint models. We compare models on information criteria, which aim to quantify the explanatory power of the models in terms of the amount of information lost, while at the same time penalizing model complexity to avoid over fitting. Specifically, we calculate an information criterion by leave-one-out cross-validation (the LOOIC, Vehtari, Gelman, & Gabry, 2017).

The "one-gait" mapping performs notably worse than the other models, achieving higher LOOIC values. That this model performs worst is a useful sanity check, because the earlier findings of worse performance in the linearly constrained model variant, and the triangular arrangement of the rates in parameter space for the unconstrained model variant should have ruled this possibility out. Next comes the "medium is special" mapping. This mapping predicted a distinct gait for medium speech, and a fall-back gait engaged to produce other speaking rates. Such a configuration might emerge as a consequence of speakers producing speech almost always around a specific habitual rate, which would become more practiced. The remaining mappings perform the best. The LOOIC estimation for the Bayesian linear switchpoint models additionally allows us to quantify the uncertainty about the point estimates of model fit. In panel E of Figure 11, lines extending from the points indicate the standard error around the LOOIC estimate. For the three best performing mappings, the standard error ranges around the point estimates are extensively overlapping, meaning that we cannot with confidence claim support for any of the three mappings ahead of the other two.

Summary of Strand 2

In Strand 2 of this study, we explored how cognitive control might be exerted on the parameters of EPONA to model speech produced at different rates. Different settings of the model parameters can be conflated with different regimes of the cognitive system underlying natural speech production. We examined how the regimes related to each other, hypothesizing that there might be "gaits" in the speech production system that speakers switch between to achieve different speaking rates. Five hypothetical mappings of rate regimes onto gaits were considered.

We found evidence that different speaking rates were achieved by distinct parameter values, and that these were arranged in a triangle in parameter space, rather than along a straight line. The triangular arrangement rules out a mapping whereby a single gait is quantitatively modulated to achieve different speaking rates. With the aim of distinguishing between the remaining mappings, we conducted further statistical modeling. This modeling ruled out one further account, namely the medium-is-special mapping, but did not allow us to distinguish between the three remaining mappings. It therefore remains an open question whether slow and medium speech is achieved by one gait and fast by another (the "fast is special" account), whether slow speech is achieved by one gait and fast and medium by another (the "slow is special" account), or whether all three rates are achieved by qualitatively distinct gaits (the "three gait" account). Nevertheless, the findings of Strand 2 provide strong evidence for a model of speech production control whereby speakers shift between different gaits to achieve different speaking rates.

General Discussion

This study had two aims. In Strand 1, we sought to establish EPONA, a new model inspired by the DBS model that would predict the duration of syllables and the duration of the overlap between them, and thereby characterize the configuration of the speech production system at different speaking rates. Subordinate to this aim, we sought to explore how the temporal properties of a word could best be encoded in the frame node.

In Strand 2, we explored how cognitive control might be exerted on the parameters of EPONA to model speech produced at different rates. Different settings of the model parameters can be seen as corresponding to different regimes of the cognitive system underlying natural speech production. We sought to examine how the regimes relate to each other, hypothesizing that there might be "gaits" in the speech production system that speakers switch between to achieve different speaking rates.

Computational Model (Strand 1)

The evolutionary algorithm learned distinct parameter settings for each speaking rate for the three model variants, though the quality of the predictions made by the trained models varied. Linear regression analyses revealed significant differences in performance between the model variants, and effect size analysis allowed us to quantify the extent of the performance differences, demonstrating a distinct performance advantage for the asynchronous model variant ahead of the control and synchronous model variants.

A salient difference between the model variants is that the control and synchronous models exhibit bimodal distributions in their fitting of the overlap duration and syllable 1 duration (see Figure 9). In contrast, the asynchronous variant predicts unimodal distributions for these objectives. It is noteworthy that the modeled syllable 1 duration and overlap duration distributions resemble each other in their overall shape. In examining the duration distributions independently for a sample of the front members (a figure showing these is included the online supplemental materials), it was plain that the bimodality of the combined distribution arises because some solutions predict distributions that contribute to the first "bump" of the bimodal distribution, and others predict distributions that contribute to the second. This result suggests that the control and synchronous model variants were not successful in finding a parameter set that solved both the serial order problem and fitted the distributions of the objectives adequately.

The observed distributions for overlap duration for all three speaking rate conditions exhibited notably less spread than the observed distributions for the two syllable duration targets. None of the model variants were particularly good at predicting the spread of the overlap, instead showing excessive spread.

Although the fits achieved by the model are satisfactory for the purposes of our Strand 2 investigation, some aspects of EPONA could potentially be revised to broaden its utility. First, the model at present is only capable of producing disyllabic words. Extending the model to produce a variety of word lengths would be relatively trivial, and would potentially allow us to explore questions regarding the extent of the gestural score, that is, are whole syllables encoded, or instead smaller segmental or demi-syllabic level chunks; or larger chunks at the level of phonological words or entire intonational phrases? Second, the current implementation of EPONA produces one word at a time, and cannot capture the interactions between previous and upcoming words, and between target words and competitors in the lexicon, although there is no reason why this could not be implemented as a network of interconnected EPONA "columns." How that might work is discussed further later.

The EPONA model follows many speech production models of the 20th century by implementing a strict separation between the formulation and execution phases (e.g., Dell & O'Seaghdha, 1992; Levelt, 1989; Levelt et al., 1999; Stemberger, 1985). The execution phase of the model is also in its conception *ballistic*, meaning that once activation arrives at the formulation-execution frontier and speech articulation begins, the gestural score will be played out without regard to what happens in the formulation phase after the onset of production.

Recent work has demonstrated that formulation and execution processes are not entirely discrete. Lexical competitors have been found to influence the details of articulation of target words (e.g., Goldrick & Blumstein, 2006; McMillan & Corley, 2010), while the articulation of slip errors has been found to differ from canonical productions of the same form (e.g., "pig" erroneously produced as [bɪg] differs from canonical "big" in voice onset time; Goldrick, Keshet, Gustafson, Heller, & Needle, 2016). Relatedly, contextual predictability and frequency predict the extent to which words are reduced by shortening the word duration and eliding segments (e.g., Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Pluymaekers, Ernestus, & Baayen, 2005). That errors and contextual priorities that arise during formulation propagate into the domain of execution has been taken as evidence in favor of cascading activation, that is, partially active "competitor" units from the formulation phase activate the corresponding articulatory plans.

A fully ballistic, cascading system would require no control on the execution phase over and above the control exerted on the formulation phase. This is of course attractive, but implausible; at the very least, a mechanism is required to allow the interruption of erroneous productions (Levelt, 1983). Alternatively, it is possible that the dynamics of the planning system after the onset of articulation also influence ongoing articulation. Fink, Oppenheim, and Goldrick (2018) set out to test the assumption of a ballistic execution component, measuring response latency and word duration in sequential picture naming tasks designed to introduce semantic interference. If the production system is ballistic, effects of semantic interference on response latency (an index of planning) and word duration (an index of articulation) should be positively correlated since variation in both metrics arises from the same process. A ballistic process cannot, however, account for effects of semantic context on duration over and above the effects correlated with the effects on latency. Fink et al. (2018) found consistent coupling of articulation and planning, compatible with the ballistic account, but also some evidence of interaction effects, suggesting that ongoing planning can exert moderate influence on execution after the onset of articulation.

Although EPONA as presented here does not explicitly model for cascading activation and has no mechanism to predict the articulatory outcome of simultaneous activation of multiple articulatory plans, it contains no features that are incompatible with the cascade concept. Similarly, the model could be considered nonballistic, in that sustained activation of the syllable gestural score is required to cause articulation of the required syllable. A more elaborate model of the execution phase might predict the articulatory outcomes of simultaneous activation of competitors (for instance, in the voice onset time of stops, as investigated by Goldrick et al., 2016), and of changes in the activation dynamics of the output nodes of the formulation network after word onset.

We followed Dell et al. (1997) in favoring a simple and interpretable model that explains the underlying psychological processes of speech production at a functional level, rather than striving for any semblance of neurobiological plausibility. The predefined activation patterns that the frame node produces on each of the ports are crucial to ensuring the correct ordering of syllable units is achieved, and have a large influence on the timing of syllable production. In general, the requirements to (a) prime upcoming units, (b) activate them at the correct time, and (c) deactivate them once they have been produced is referred to as the serial order problem. Dell, Burger, and Svec's (1997) approach to resolving the serial order problem using predefined activation patterns is functional and minimal.

It is, however, also possible to achieve correct serial ordering using only components from the standard connectionist toolbox. In this respect, a promising approach is competitive queueing (Grossberg, 1978; Houghton, 1990), which employs a two-layer subnetwork to maintain serial order. The first layer is a planning layer, where all nodes for all the elements in a sequence become active in parallel, with their relative activation encoding the order of realization (a primacy gradient). The nodes of the planning layer project onto the same number of nodes in the second, competitive choice layer, where inhibitory connections ensure that only the activation of the most active node at any given time is transmitted to the output nodes, and a switch-off mechanism ensures that successfully produced items are inhibited, allowing subsequent items to be produced (see Hurlstone et al., 2014, for an extensive review). It would be fruitful to evaluate a model that employed competitive queueing in the frame node. This would remove the need for the implausible stepped activation patterns in the frame node.

The activation function (that is, the function that computes the activation of a node from the activation arriving at it through connections, also known as a transfer function) in the model is strictly linear. Sigmoid activation functions such as tanh (Harm & Seidenberg, 1999) or soft-max (Chang, 2002; Chang, Dell, & Bock, 2006) are employed in several more recent models where competition between nodes at the same level is modeled. It is, however, unlikely that a different choice of activation function would have made a large difference to the outcomes of this study, because our model does not simulate between-node competition. In a model with competitive queuing, a nonlinear activation function might prove advantageous.

How Do Regimes Relate to Each Other? (Strand 2)

Because the asynchronous model variant performed significantly better than either the control or synchronous model variants, we performed analyses in parameter space only for this variant. The following discussion refers therefore to the asynchronous model variant only.

The speaking rate regimes identified in this investigation can be compared along two dimensions; first, in terms of the parameter values that the model engages to achieve each targeted speaking rate (comparison in parameter space), and second in terms of the predicted fingerprint durations (comparison in prediction space).

Which, if any, gaits are present? To distinguish between single-gait and multiple-gait scenarios, we examined the arrangement of the regimes in parameter space. We predicted that in the single-gait scenario, the three regimes would be arranged sequentially along an axis in parameter space. In a multiple gait scenario, the three regimes would be arranged in a triangle in parameter space. The arrangement of the optima on the PC1-PC2 plane was clearly nonaxial (see Figure 10). Our results therefore indicate that cognitive regimes adopted to achieve different speaking rates are arranged in a manner that is incompatible with a single-gait system.

It could however, still be the case that, although the optimization routine had settled on a nonlinear arrangement of rates, a linear arrangement could have been able to fit the data adequately. A further asynchronous model variant was fitted to test this, where the arrangement of the rates in parameter space was constrained to be linear or axial. This model fitted the data less well than the unconstrained model, reinforcing our conclusion that multiple gaits are present.

Having established that the single gait configuration was unlikely given the data, we moved on to comparing the regimes in prediction space. Aside from all rates being produced by one gait, there are four further possible mappings of rates onto gaits: three gaits; slow is distinct while fast and medium are mapped to the same gait; fast is distinct; medium is distinct.

The plausibility of these mappings could be teased apart by examining the extent of nonlinearity in the predicted distributions of models fitted with parameter values taken from the spaces *between* the centers identified in the evolutionary optimization. We performed statistical fitting to test for (non)linearity along the axes linking the center points of each rate, and compared the quality of fit of models instantiating the five possible mappings. We used Bayesian linear switchpoint models, which are able to fit variation in the spread of the distribution, and allowed us to quantify certainty at all stages of modeling, including model comparison.

This statistical modeling allowed us to directly test the plausibility of the five mappings. The one-gait mapping was rejected, consistent with the triangular arrangement of the rates in parameter space and the rejection of the model variant with the linearity constraint in the optimization paradigm. Support for the “medium is special” mapping was limited. Although the “three-gaits” mapping had numerically the best fit, the statistical modeling was unable to distinguish between this mapping and the “fast is special” and “slow is special” accounts. This means that all three mappings are plausible models of the cognitive reality, given the present dataset and modeling approach. While we believe that the statistical modeling is sufficiently sensitive to evaluate the plausibility of the mappings, it is of course dependent on the data provided by the simulations. These data may be insufficient in two ways. First, they consist only of predicted distributions of the three fingerprint durations, which may not be rich enough a representation of the acoustic reality to highlight subtle differences in linearity between the speaking rates. Second, the variability that was valuable in the parameter optimization paradigm for the reconstruction of the distributions to be compared with the observed distributions may have proved counterproductive for the statistical modeling we conducted.

Further experimental work is required to clarify the nature of the mapping of speaking rates to gaits, possibly testing more than three speaking rates in a denser sampling.

The consequences of the presence of gaits for models of speech production and perception. Our concept of different “gaits,” each encompassing qualitatively similar regimes in the formulation component of the speech production system, represents a theoretical step forward that makes predictions that may be fruitfully explored in future modeling and empirical work, building on the conception of gaitedness at the execution level.

Although this study concerned speaking rate variation and demonstrated the presence of cognitive gaits to achieve different speaking rates, it is plausible to think of shifting between qualitatively different parameter regimes as a more general mechanism to deliberately modulate the acoustic and temporal properties of speech to suit various communicative situations (Lindblom, 1990; Lindblom, Brownlee, Davis, & Moon, 1992).

Natural speech produced by any one speaker varies in many more ways than along a single dimension of speaking rate, in effect adopting what has often been called different *registers* or *speaking styles* (Hirschberg, 2000). It has been observed that speakers transform the acoustics of their speech to enhance its intelligibility for their interlocutor, or in response to the reverberance or background noise of their environment (Cooke et al., 2014). Prepared speech, such as reading aloud, varies from spontaneous speech (e.g., Furui, 2003). Typically, these speaking styles have been thought of (or at least treated as) categorically distinct, driven perhaps by the methodologies used to elicit the speech during experiments and corpus gathering, or to categorize the situations in which the speech arose in generalist corpora (Hirschberg, 2000).

Although acoustic differences emerge between speech categorized according to these situational categories, knowing that such differences exist says little about how speakers modulate the speech formulation and execution mechanisms to achieve that variation. This is because it remains unknown to what extent the speech planning system engages categorically distinct regimes to achieve different speaking styles, and whether these researcher-imposed situational labels bear any resemblance to the underlying cognitive categories.

If different speaking styles are achieved by switching between qualitatively different gaits of the speech planning system, we would expect there to be observable clustering in the acoustic characteristics of speech across the range of speech variability, reflecting the categorical shifts between cognitive gaits. Two recent findings suggest that speaking style variation may be at least to some extent categorical. The first concerns reduced pronunciation variants, that is, pronunciations of words where acoustic cues, segments, and sometimes entire syllables are omitted, generally when words are highly predictable and in informal spontaneous speaking situations (e.g., Ernestus, Hanique, & Verboom, 2015; Ernestus & Warner, 2011), for example the realization of American English "yesterday," the canonical form of which is /jɛstɚɾeɪ/, as [jɛʃeɪ]. Reduction of this type is one of the ways in which acoustic differences between speaking styles surface and can be quantified. Hanique, Ernestus, and Schuppler (2013) found evidence that both categorical and gradient processes were simultaneously responsible for an instance of schwa deletion in Dutch.

The second concerns the retrieval of speaking style labels through machine-learning techniques. Bentum, Ernestus, ten Bosch, and van den Bosch (2019) employed a language modeling and dimensionality reduction approach to characterize word choice and co-occurrence across the speaking styles in the orthographic transcriptions of a corpus of Dutch speech containing many different speaking styles (Oostdijk, 2000). Many of the speaking styles labeled in the corpus emerged as distinct clusters, while other groups of speaking styles merged to form a single cluster. Again, this hints that, underlyingly, speaking styles differ categorically from each other on various dimensions.

The finding of gaitedness in speech production has consequences for models of speech perception. If the speech produced by speakers varies qualitatively between gaits, then listeners might also be expected to adopt different processing strategies to make the most of the cues available in the speech signal associated with a specific gait. If that were the case, we might expect to see gaits in speech perception to mirror those in speech production.

Extending EPONA to a network model. This research has not addressed how gaits could be manifested in the structure of the lexicon, instead focusing on exploring the parameter settings in a single column of the EPONA model that must be altered to achieve different speaking rates. In this section, we explore how the EPONA model could be extended to form a network model of the lexicon. An EPONA network model would facilitate exploration of how gait selection would work as a mechanism to control speech rate in multiword, continuous speech.

A network view of EPONA, as illustrated in Figure 12 isolates the different parameter settings associated with each gait in a "variant" frame node for the relevant word shape, which is in turn connected to "variant" structure nodes encoding different temporal realizations of the relevant structure. Each word in the lexicon is connected to all frame nodes suitable to produce that wordshape.

We postulate exhaustive excitatory and inhibitory connections between related variants of different frame nodes, and inhibitory connections between variants of the same frame node. This interconnection causes priming activation and suppression that tends to ensure that adjacent words are produced at with the same speaking style. We will call these connections between the frame nodes the "reinforcement route," consistent with the segmental and metrical routes.

In the reinforcement route, networks or families of related frame nodes, depicted in Figure 12 as different colors of nodes, are connected together by heavily weighted connections. Although three families of frame nodes are depicted in Figure 12 for each word form, it is clear that different word shapes will have differing number of frame node variants, reflecting different possibilities for categorical reduction. Some frame node variants will therefore belong to multiple interconnected networks of related frame nodes. A gait in the EPONA model is then the reinforcement network of strongly interconnected frame node variants, which tend to prime each other, and whose priming activation tends to suppress the frame node variants belonging to other gaits. The weightings of the connections in

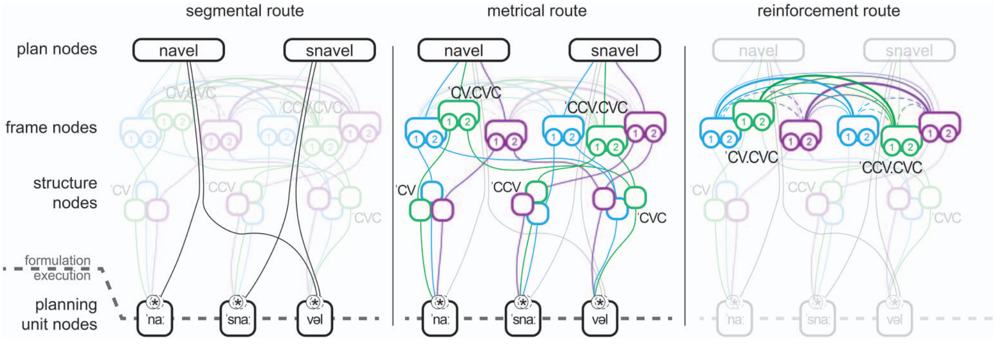

Figure 12. A sketch of an EPONA network containing the nodes necessary to produce the Dutch disyllabic words *navel* [ˈnaː.vəl] "navel" and *snavel* [ˈsnaː.vəl] "beak," adding the reinforcement route and multiple frame and structure nodes to capture gaited behavior. See the text for details. See the online article for the color version of this figure.

the reinforcement networks are learnt by associative learning, such that frame nodes that co-occur in time develop strong excitatory connections, and those that do not co-occur develop inhibitory connections.

These gait reinforcement networks are also the site of influence from higher processes that modulate the selection of production variants. For instance, if the speaker is in a conversational situation that invites slow, canonical production, then they will engage executive control functions to inhibit gait networks that contain frame nodes for fast, highly reduced variants. Other, less explicit, contextual factors could be thought to have a similar mechanism. For instance, a "prosodic moderator" might excite gait networks containing lengthened productions preceding prosodic phrase boundaries, or during the preparation of words that should be marked as extra prominent (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992).

The excitatory connections within gait networks and inhibitory connections between them have the property of introducing inertia into the system: The default behavior is always to continue speaking in the same gait. This means that switching gait should be effortful, because it requires invocation of executive control to suppress activation in competing gait reinforcement networks and to boost activation in the target gait reinforcement network. The effectiveness and the speed of the switch should also be related.

If executive control is invoked to moderate the activation in competing and target gait networks, it might be possible to detect correlations between the various components of executive control ability (Miyake et al., 2000) and individual differences between speakers in their success at modulating their speaking rate. Relatedly, the modeling techniques developed for this study could also be used to explore individual differences between speakers in the cognitive regimes that they invoke to produce different speaking rates.

The gait for formulation that we have described in this article resemble locomotive gaits in that they are qualitatively distinct configurations that can be invoked to adjust the speed of a behavior. However, they do not capture a striking feature of locomotive gaits, namely that gaits are selected based on their biomechanical efficiency to achieve the required movement speed (Hoyt & Taylor, 1981). How the biomechanical effort differences inherent to locomotion gaits might be mirrored in analogous speaking gaits is clear for the execution component, where movement distance and precision are obvious candidates. In a network view of the EPONA model, effortfullness would be the sum of two components: the degree of executive control engagement needed to maintain the target gait, and the difficulty of performing conceptual retrieval and phrase level planning fast enough to keep up with the formulation and execution components.

Conclusion

We proposed that to achieve different speaking rates, the speech planning system adopts different configurations, or regimes. Because speakers are able to voluntarily adjust their speaking rate, they must have a control mechanism that enables them to shift from regime to regime. Describing the way in which these regimes are arranged relative to each other in parameter space is highly informative for understanding the nature of the control mechanism that is engaged to shift between regimes, and how control might be exerted on speech production in general. We hypothesized that speech rate control might be achieved by shifting between different, qualitatively distinct "gaits" of the speech production mechanism. Alternatively, different speaking rates might be achieved by continuous adjustment within a single rate.

We set ourselves the task of distinguishing these hypotheses. We developed EPONA, a model inspired by the influential DBS model (Dell et al., 1997), to predict the distributions of syllable and syllable-overlap durations that characterize speech production in a specific speaking rate regime. By optimizing the parameters of this model to fit each of three rate conditions independently, we identified optimal parameter settings for each speaking rate, which we conflate with the dimensions of the regime-space of the underlying cognitive system. By examining the arrangement of the parameter optima of the model, we could infer the arrangement of the underlying cognitive system. The model optima resembled a triangle (see Figure 10), rejecting the idea that the regimes of the speech production system all belong to a single qualitatively consistent gait. By fitting further models where linearity in parameter space was enforced, we provided further evidence ruling out a single-gait account.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

Alexander, R. M. (1989). Optimization and gaits in the locomotion of vertebrates. *Physiological Reviews, 69,* 1199–1227. http://dx.doi.org/10.1152/physrev.1989.69.4.1199

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60,* 92–111. http://dx.doi.org/10.1016/j.jml.2008.06.003

Bella, S. D., & Palmer, C. (2011). Rate effects on timing, key velocity, and finger kinematics in piano performance. *PLoS ONE, 6,* e20518. http://dx.doi.org/10.1371/journal.pone.0020518

Bentum, M., Ernestus, M., ten Bosch, L., & van den Bosch, A. (2019). Do speech registers differ in the predictability of words? *International Journal of Corpus Linguistics, 24,* 98–130.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological review, 89,* 1–47.

Booij, G. (1995). *The phonology of Dutch*. Oxford, UK: Clarendon Press.

Boomer, D. S., & Laver, J. D. M. (1968). Slips of the tongue. *British Journal of Disorders of Communication, 3,* 2–12. http://dx.doi.org/10.3109/13682826809011435

Bosker, H. R., & Cooke, M. (2018). Talkers produce more pronounced amplitude modulations when speaking in noise. *The Journal of the Acoustical Society of America, 143,* EL121–EL126.

Boyce, S., & Espy-Wilson, C. Y. (1997). Coarticulatory stability in American English /r/. *The Journal of the Acoustical Society of America, 101,* 3741–3753. http://dx.doi.org/10.1121/1.418333

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

Cambier-Langeveld, T., Nespor, M., & van Heuven, V. J. (1997). The domain of final lengthening in production and perception in Dutch. *EUROSPEECH-1997* (pp. 931–934). Retrieved from https://www.isca-speech.org/archive/eurospeech_1997/e97_0931.html

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science, 26,* 609–651.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review, 113,* 234–272.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83,* 596–610.

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language, 28,* 543–571.

Damian, M. F., & Dumay, N. (2007). Time pressure and phonological advance planning in spoken production. *Journal of Memory and Language, 57,* 195–209. http://dx.doi.org/10.1016/j.jml.2006.11.001

Deb, K., & Agrawal, R. B. (1995). Simulated binary crossover for continuous search space. *Complex Systems, 9,* 115–148.

Deb, K., & Goyal, M. (1996). A combined genetic adaptive search (GeneAS) for engineering design. *Computer Science and Informatics, 26,* 30–45.

Deb, K., & Jain, H. (2014). An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, Part I: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation, 18,* 577–601. http://dx.doi.org/10.1109/TEVC.2013.2281535

Deb, K., Sindhya, K., & Okabe, T. (2007). Self-adaptive simulated binary crossover for real-parameter optimization. *Proceedings of the 9th annual conference on genetic and evolutionary computation* (pp. 1187–1194). New York, NY: ACM. http://dx.doi.org/10.1145/1276958.1277190

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93,* 283–321.

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104,* 123–147.

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition, 42,* 287–314. http://dx.doi.org/10.1016/0010-0277(92)90046-K

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21,* 1664–1670. http://dx.doi.org/10.1177/0956797610384743

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112,* 777–813. http://dx.doi.org/10.1037/0033-295X.112.4.777

Ernestus, M., Hanique, I., & Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *Journal of Phonetics, 48,* 60–75. http://dx.doi.org/10.1016/j.wocn.2014.08.001

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics, 39,* 253–260. http://dx.doi.org/10.1016/S0095-4470(11)00055-6

Fink, A., Oppenheim, G. M., & Goldrick, M. (2018). Interactions between lexical access and articulation. *Language, Cognition and Neuroscience, 33,* 12–24. http://dx.doi.org/10.1080/23273798.2017.1348529

Furui, S. (2003). Recent advances in spontaneous speech recognition and understanding. *ISCA & IEEE workshop on spontaneous speech processing and recognition*. Retrieved from https://www.isca-speech.org/archive_open/sspr2003/sspr_mmo1.html

Garnham, A., Shillcock, R. C., Brown, G. D., Mill, A. I., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics, 19,* 805–818.

Garrett, M. (1976). Syntactic processes in language production. In R. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 231–256). Amsterdam, the Netherlands: North Holland Publishing Company.

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21,* 649–683. http://dx.doi.org/10.1080/01690960500181332

Goldrick, M., Keshet, J., Gustafson, E., Heller, J., & Needle, J. (2016). Automatic analysis of slips of the tongue: Insights into the cognitive architecture of speech production. *Cognition, 149,* 31–39. http://dx.doi.org/10.1016/j.cognition.2016.01.002

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—A syllable-centric perspective. *Journal of Phonetics, 31,* 465–485.

Grossberg, S. (1978). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology, 17,* 199–219.

Guenther, F. H. (2016). *Neural control of speech*. Cambridge, MA: MIT Press.

Hadka, D. (2017). *Platypus: A free and open source Python library for multiobjective optimization*. Project Platypus. Retrieved from https://github.com/Project-Platypus/Platypus

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In V. Gael, M. Jarrod, & V. Travis (Eds.), *Proceedings of the 7th Python in Science Conference (SciPy2008)* (pp. 11–15). Retrieved from https://conference.scipy.org/proceedings/scipy2008/paper_2/

Hanique, I., Ernestus, M., & Schuppler, B. (2013). Informal speech processes can be categorical in nature, even if they affect many different words. *The Journal of the Acoustical Society of America, 133,* 1644–1655.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106,* 491–528.

Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. In H. Merle (Ed.), *Prosody: Theory and experiment* (pp. 335–350). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In M. Zock & R. Dale (Eds.), *Current research in natural language generation* (pp. 287–319). London, UK: Academic Press.

Hoyt, D. F., & Taylor, C. R. (1981). Gait and the energetics of locomotion in horses. *Nature, 292,* 239–240. http://dx.doi.org/10.1038/292239a0

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin, 140,* 339–373.

Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., & Bosker, H. R. (2019). Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000744

Kello, C. T., Plaut, D. C., & MacWhinney, B. (2000). The task dependence of staged versus cascaded processing: An empirical and computational study of Stroop interference in speech production. *Journal of Experimental Psychology: General, 129,* 340–360.

Kelso, J. A., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14,* 29–59.

Kennedy, J. (2011). Particle swarm optimization. In *Encyclopedia of machine learning* (pp. 760–766). http://dx.doi.org/10.1007/978-0-387-30164-8_630

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14,* 41–104. http://dx.doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences, 22,* 1–38.

Lindblom, B. (1990). Explaining phonetic variation: A Sketch of the H &

H. theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). http://dx.doi.org/10.1007/978-94-009-2037-8_16

Lindblom, B., Brownlee, S., Davis, B., & Moon, S.-J. (1992). Speech transforms. *Speech Communication, 11,* 357–368. http://dx.doi.org/10.1016/0167-6393(92)90041-5

Lindblom, B., Lubker, J., & Gay, T. (1977). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of The Acoustic Society of America, 62,* S15.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia, 8,* 323–350. http://dx.doi.org/10.1016/0028-3932(70)90078-3

MacKay, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology, 3,* 210–227. http://dx.doi.org/10.1016/0010-0285(72)90004-7

Martí, L., Garcia, J., Berlanga, A., & Molina, J. M. (2009). An approach to stopping criteria for multi-objective optimization evolutionary algorithms: The MGBM criterion. *2009 IEEE Congress on Evolutionary Computation* (pp. 1263–1270). http://dx.doi.org/10.1109/CEC.2009.4983090

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). How the tracking of habitual rate influences speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45,* 128–138.

McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition, 117,* 243–260. http://dx.doi.org/10.1016/j.cognition.2010.08.019

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica, 41,* 215–225. http://dx.doi.org/10.1159/000261728

Minetti, A. E. (1998). The biomechanics of skipping gaits: A third locomotion paradigm? *Proceedings of the Royal Society B: Biological Sciences, 265,* 1227–1235.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100.

Oostdijk, N. H. J. (2000). Het corpus gesproken Nederlands [The spoken Dutch corpus]. *Nederlandse Taalkunde, 5,* 280–284.

Parrell, B. (2012). The role of gestural phasing in Western Andalusian Spanish aspiration. *Journal of Phonetics, 40,* 37–45. http://dx.doi.org/10.1016/j.wocn.2011.08.004

Peer, E. S., Bergh, F. v. d., & Engelbrecht, A. P. (2003). Using neighbourhoods with the guaranteed convergence PSO. *Proceedings of the 2003 IEEE Swarm Intelligence Symposium, 2003. SIS '03* (pp. 235–242). http://dx.doi.org/10.1109/SIS.2003.1202274

Pennycuick, C. J. (1975). On the running of the gnu (Connochaetes taurinus) and other animals. *Journal of Experimental Biology, 63,* 775–799.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica, 62,* 146–159. http://dx.doi.org/10.1159/000090095

Pouplier, M. (2012). The gaits of speech. In M.-J. Solé & D. Recasens i Vives (Eds.), *The initiation of sound change: Perception, production, and social factors* (Vol. 323, pp. 147–164). Amsterdam, the Netherlands: John Benjamins Publishing.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America, 123,* 1104–1113.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rodd, J., Bosker, H. R., ten Bosch, L., & Ernestus, M. (2019). Deriving the onset and offset times of planning units from acoustic and articulatory measurements. *The Journal of the Acoustical Society of America, 145,* EL161–EL167. http://dx.doi.org/10.1121/1.5089456

Rodd, J., Bosker, H. R., ten Bosch, L., Ernestus, M., & Meyer, A. S. (2019). *PiNCeR: A corpus of cued-rate multiple picture naming in Dutch*. Unpublished manuscript. http://dx.doi.org/10.31234/osf.io/wyc6h

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language, 162,* 42–45.

Schiel, F. (2015). A statistical model for predicting pronunciation. In M. Wolters, J. Livingstone, B. Beattie, J. Stuart-Smith, & J. Scobbie (Eds.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow.

Scobbie, J. M., & Pouplier, M. (2010). The role of syllable structure in external sandhi: An EPG study of vocalisation and retraction in word-final English /l/. *Journal of Phonetics, 38,* 240–259. http://dx.doi.org/10.1016/j.wocn.2009.10.005

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In C. E. William & W. C. T. Edward (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295–342). Hillsdale, NJ: Lawrence Erlbaum Associates.

Slootweg, A. (1988). Metrical prominence and syllable duration. In C. Peter & H. Aafke (Eds.), *Linguistics in the Netherlands* (pp. 139–138). Dordrecht, the Netherlands: Foris Publications.

Stan Development Team. (2018). RStan: The R interface to Stan (R package version 2.18.2) [Computer software]. Retrieved from http://mc-stan.org/

Stemberger, J. P. (1985). An interactive activation model of language production. In W. E. Andrew (Ed.), *Progress in the psychology of language* (Vol. 1, pp. 143–186). Hillsdale, NJ: Erlbaum.

Stemberger, J. P. (1991). Apparent anti-frequency effects in language production: The addition bias and phonological underspecification. *Journal of Memory and Language, 30,* 161–185. http://dx.doi.org/10.1016/0749-596X(91)90002-2

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26,* 952–981. http://dx.doi.org/10.1080/01690960903498424

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27,* 1413–1432. http://dx.doi.org/10.1007/s11222-016-9696-4

Vousden, J. I., Brown, G. D., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology, 41,* 101–175.

Vousden, J. I., & Maylor, E. A. (2006). Speech errors across the lifespan. *Language and Cognitive Processes, 21,* 48–77. http://dx.doi.org/10.1080/01690960400001838

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review, 114,* 830–841.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review, 120,* 329–355. http://dx.doi.org/10.1037/a0031542

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America, 91,* 1707–1717.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society Series C, 22,* 392–399. http://dx.doi.org/10.2307/2346786

Wood, S. N. (2017). *Generalized additive models: An Introduction with R* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Wright, C. E., & Meyer, D. E. (1983). Conditions for a linear speed-accuracy trade-off in aimed movements. *The Quarterly Journal of Experimental Psychology Section A, 35,* 279–296. http://dx.doi.org/10.1080/14640748308402134

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., . . . Woodland, P. (2006). *The HTK book (for HTK version 3.4)*. Cambridge, UK: Cambridge University Engineering Department.

Zitzler, E., Brockhoff, D., & Thiele, L. (2007). The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, & T. Murata (Eds.), *Evolutionary multi-criterion optimization* (pp. 862–876). Berlin, Germany: Springer.

Received August 28, 2018
Revision received September 9, 2019
Accepted September 10, 2019 ■