HHS Public Access

Author manuscript

J Child Lang. Author manuscript; available in PMC 2024 November 01.

Published in final edited form as:

J Child Lang. 2023 November ; 50(6): 1318–1335. doi:10.1017/S0305000923000260.

Neurocomputational modeling of speech motor development

Andrew M. Meier[1], **Frank H. Guenther**[1,2]

[1]Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA 02215

[2]Department of Biomedical Engineering, Boston University, Boston, MA 02215

Abstract

This review describes a computational approach for modeling the development of speech motor control in infants. We address the development of two levels of control: articulation of individual speech sounds (defined here as phonemes, syllables, or words for which there is an optimized motor program) and production of sound sequences such as phrases or sentences. We describe the DIVA model of speech motor control and its application to the problem of learning individual sounds in the infant's native language. Then we describe the GODIVA model, an extension of DIVA, and how chunking of frequently produced phoneme sequences is implemented within it.

Keywords

speech motor control; computational neural modeling; speech development; motor-sequence learning; speech production

I. The DIVA model of speech motor control

The Directions Into Velocities of Articulators (DIVA) model is an artificial neural network that provides a quantitative account of the computations underlying speech motor control (Guenther 1995; Tourville and Guenther 2011; E. Golfinopoulos, Tourville, and Guenther 2010; see Guenther, 2016 for a detailed treatment). It contains a network of simulated components which represent brain structures responsible for producing speech. The model includes an articulatory synthesizer that mimics the behavior of the vocal tract, and the neural network learns to control movements of the synthesizer's articulators in order to produce intelligible speech. We focus herein on a higher-level treatment of the model's neural computations and developmental processes, avoiding mathematical equations and computer implementation details for tractability.

To understand the model, we will start by defining a *speech sound* to be a "chunk" of speech that has its own optimized motor program in the brain. These chunks could be phonemes, syllables, and/or words, depending on the age and linguistic experience being considered. In keeping with a number of prior proposals (e.g., Kozhevnikov and Chistovich 1965; Levelt 1993; MacNeilage and Davis 1990) and supported by distributional analyses of

Correspondence concerning this article should be addressed to Andrew Meier (amsmeier@bu.edu).

No competing interests to disclose.

phoneme combinations (Sun and Poeppel 2022; Kessler and Treiman 1997), we suggest that the syllable is the most typical sound chunk with an optimized motor program. However, motor programs likely also exist for individual phonemes as well as frequently produced multisyllabic utterances, such as common words or names of familiar people and locations. Note that the motor programs can be hierarchical; for example, a syllabic motor program will consist of individual phoneme motor programs along with optimized transitions between these phoneme motor programs.

The model assumes that, in the mature speaker, speech production begins with an intended linguistic message being translated by higher-level brain regions into a sequence of speech sounds. Motor sequencing circuits then activate the appropriate nodes of a *speech sound map* in ventral premotor cortex (vPMC), which is the highest processing level represented in DIVA. While this model focuses on segmental control - production of phonemes, syllables, and words - it should be noted that prosodic control is also an essential goal of speech motor development (Mattys et al. 1999; Kehoe and Stoel-Gammon 1997).

Neural components of the DIVA model

The brain structures whose functions are simulated by the DIVA model are illustrated in Figure 1. Each box corresponds to a set of modeled neurons, or *nodes*, that together form a neural map of some type of speech-relevant information. Larger boxes indicate cortical regions and smaller boxes indicate subcortical nuclei. Arrows represent excitatory projections while circles represent inhibitory projections, with the projection target being the area touching the arrowhead or circle. Production of a speech sound starts with activation of a node representing that particular sound in a *speech sound map* in the left ventral premotor cortex. Activation of this node leads to motor commands that arrive in motor cortex via two control systems: a *feedforward control system* and a *feedback control system*.

The feedforward control system generates previously learned motor programs for speech sounds. This process involves two components. The first component of feedforward control ensures that the motor program is initiated at the appropriate time. Timing control is carried out by a cortico-basal ganglia loop that includes an *initiation map* in the supplementary motor area (SMA). This loop identifies the appropriate sensory, motor, and cognitive context for producing the speech sound. We suggest that the input structures of the basal ganglia monitor these contextual cues, with the caudate monitoring cognitive context and the putamen monitoring sensory and motor contexts. When the appropriate context for producing a speech chunk is identified, a corresponding node is activated in the initiation map via the globus pallidus (GP), substantia nigra pars reticula (SNr), and the ventral anterior (VA) thalamic nucleus. This initiation map node activation triggers the readout (execution) of the learned motor program for the current speech sound.

The second component of the feedforward control system comprises the motor programs themselves, which generate feedforward commands for producing learned speech sounds. These commands are encoded by synaptic projections from the speech sound map to an articulator map in the right and left ventral primary motor cortex (vMC). The cortico-cortical projections from left vPMC to vMC are supplemented by a cerebellar loop passing through the pons, cerebellar cortex lobule VI (Cb-VI), and the ventral lateral (VL) nucleus of

thalamus. This division of motor execution between cerebellar and basal ganglia loops was originally proposed in a theory founded on nonhuman primate neurophysiology (Hikosaka et al. 2002), with later support being provided by human neuroimaging (Doyon et al. 2009). Note that multiple instances of a structure in Figure 1, such as the Cb, are implemented as separate non-overlapping neural populations within that structure. For example, separate Cb networks process feedforward commands, auditory targets, and somatosensory targets.

The *auditory feedback control subsystem* detects and corrects for mismatches between the auditory target and the current auditory feedback. Axonal projections from speech sound map nodes in vPMC - both directly and via a cortico-cerebellar loop involving the pons, cerebellum (Cb), and medial geniculate (MG) nucleus of the thalamus – arrive at an *auditory target map* in the higher-order auditory cortical areas in posterior auditory cortex (pAC), including the posterior superior temporal gyrus and sulcus and the planum temporale. These projections signal the expected auditory percept generated by the sound currently being produced.

The auditory target for the current sound is compared to incoming reafferent auditory signals. This information is transmitted to cortical areas via MG and is represented in the model's *auditory state map*. If the current auditory state does not match the target, auditory error nodes in the higher-order auditory cortical areas become active. These types of predictive and error-related responses have been localized to auditory cortex by neural recordings in humans (Hashimoto et al. 2003; Hickock et al. 2018; Ozker et al. 2022). Auditory error node activities are then transformed into corrective motor commands through projections from the auditory error nodes to the *feedback control map* in right vPMC, which in turn projects to the articulator map in vMC both directly and via a loop through the pons, Cb, and VL. Auditory error is computed as a simple subtraction of the target from the state. This subtraction is enabled by making the Auditory State, Target, and Error Maps contain identical representations of speech sounds and equalizing the strength of inputs from the Target and State Maps to the Error Map.

The DIVA model also contains a *somatosensory feedback control subsystem,* the main components of which are hypothesized to reside in ventral somatosensory cortex (vSC). Projections from the speech sound map to the *somatosensory target map* encode the expected somatosensory feedback during sound production. These projections include cortico-cortical as well as cortico-cerebellar loop projections via the ventral posterior medial (VPM) thalamic nucleus. The model's *somatosensory state map* represents proprioceptive and tactile information from the speech articulators. If the somatosensory state does not match the current target, the *somatosensory error map* sends a corrective command via the feedback control map to correct subsequent motor commands. Studies in which articulator sensory feedback is perturbed during speaking suggest that the somatosensory error map resides primarily in ventral somatosensory cortex (Golfinopoulos et al. 2011).

The components of the DIVA model are a set of heterogeneous, biophysically realistic neural networks. Different neural network structures were chosen for each component based on the distinct function they serve. For example, different architectures were required for the error maps, which compute differences between two input signals, and the Initiation

Map, which controls the timing of activation in a downstream structure. Some components in Figure 1 were not instantiated as full neural networks, such as VA and VL, which serve as simple relays from the basal ganglia to the cortex.

Unlike other models of speech motor control (e.g. Hickock 2014), feedforward commands in DIVA proceed directly to primary motor cortex, without comparison to an internal model of sensory consequences. The lack of sensorimotor knowledge present at this processing stage is not problematic in the scenarios addressed by the model, in which auditory targets have already been well learned. However, this simplification does reduce the application of DIVA in particular speech phenomena, such as internal error correction (Nozari et al. 2011) and attempting to imitate unfamiliar sounds (e.g. Hao and Jong 2016).

Because most projections in the model are long-range and originate in the cerebral cortex, they are modeled as excitatory, to match known neuroanatomy (DeFelipe and Farinas 1992; see Urrutia-Pinones et al. 2022 regarding exceptions to this pattern). In the case of error maps, inputs are modeled as inhibitory, which is necessary for detecting differences between sensory states and sensory targets. Correlates in the brain of these projections to error maps likely use feedforward inhibition, in which a source area provides long-range excitatory projections to inhibitory neurons in a target area, effectively inhibiting certain excitatory neurons in that target area (Li et al. 2014; Naskar et al. 2021). All pathways in Figure 1 are assumed to have been established by birth, though the micro-scale patterns and weights of connections maintain plasticity, allowing for further postnatal development (Kostović and Jovanov-Milošević 2006; Dubois et al. 2014).

Implementation of speech motor learning in DIVA

In order for the DIVA model to produce speech, it must undergo a learning process analogous to what occurs in the developing infant brain. The stages of this process are simplified for the purposes of implementation into a *babbling phase* and an *imitation phase*.

The babbling phase involves the generation of semi-random articulator movements through activation of nodes in the model's articulation map (corresponding to vMC), which drives movements of the speech articulators and the generation of auditory and somatosensory feedback signals. The resulting combination of auditory, somatosensory, and articulatory representations is used to tune inverse models that map somatosensory and auditory errors into corrective motor commands via the *feedback control map* in Figure 1. The learning in this stage is not phoneme- or syllable-specific; the learned sensory-motor transformations are applicable to all speech sounds that will be learned later.

During the imitation phase, the model is presented with sample speech sounds to learn, similar to an infant being exposed to the sounds of their native language. These sounds take the form of time-varying acoustic signals corresponding to phonemes, syllables, or words. Based on these samples, the model first learns an auditory target for each sound. Learning of a sound's auditory target involves activation of a speech sound map node that will later represent the sound for production. This occurs via a speech recognition system when the model "hears" the sound[1], which corresponds to a child hearing a new speech sound directed at him/her them by a parent, for example. This in turn leads to adjusting

synaptic weights in the projections from that speech sound map node to the auditory cortex to encode the sound's auditory target.

After an auditory target for a sound has been learned, the model can attempt to produce the sound. The appropriate nodes in the initiation map and speech sound map must first be activated. At first, the model will not have a tuned motor program for producing the sound in a feedforward manner, nor will it have a somatosensory target. Thus, the system will depend primarily on auditory feedback for guidance. On each production attempt, the motor target will be updated to incorporate the commands generated by the auditory feedback control subsystem on that attempt. These commands are generated by first determining the auditory error (i.e., the distance and direction in auditory space between the target and what was produced) in the Auditory Error Map. The auditory error is then sent to the Feedback Control Map, where it is transformed into articulator movements that will reverse the auditory error. This corrective signal is then sent to the Articulator Map, where it adjusts the velocities of articulator movements. Subsequent attempts will then have a more accurate feedforward command to guide production.

Over time, the feedforward commands will become sufficient by themselves for reliably producing the sound. That is, the motor program will have become accurate enough that it generates very few auditory errors, obviating the need for auditory feedback control in most instances. At this point the model can fluently produce the speech sound. As the speech articulators grow, the auditory feedback control subsystem continually corrects for changes in the biomechanics of the vocal tract. These corrective commands are subsumed into the motor program, thus allowing it to stay tuned despite significant changes to the shapes and sizes of the articulators over the course of life.

As the model repeatedly produces a sound, it also learns a somatosensory target region for that sound, analogous to the auditory target region. The somatosensory target represents the expected proprioceptive and tactile sensations elicited when producing the sound. This target is different from the auditory target in that it cannot be learned from other speakers, as essential information about tactile patterns, tongue shape, etc. are not available to a listener. The somatosensory target must instead be learned through self-monitoring of one's own correct productions, a process that occurs at a later stage than the learning of auditory targets.

The simulation study of Callan et al. 2000 provides an example of how the DIVA model has been used to investigate speech motor development. This study involved computer simulations of the process of learning and correctly producing English vowels during developmental growth of the vocal tract. The model was grounded in empirical data by including the sizes and shapes of infant vocal tracts measured with magnetic resonance imaging. Vowel formants were successfully produced along a developmental timeline that matched those observed in real developing infants, showing the feasibility of the model. The simulation provided additional insight into speech development by showing how infants

[1]In model simulations, the speech recognition system is not implemented; instead, sound identity is provided by the modeler, who labels the speech sounds presented to the model for learning.

could make use of motor equivalence to produce a sound, even under the constraints of changing articulator shapes and sizes.

II. Development of speech motor programs

The motor learning process implemented in computer simulations of the DIVA model as described in the previous section is a highly simplified approximation of speech development in children. In the current section, we provide a more detailed account of the stages of speech development in infants and children with reference to components of the DIVA model.

Overview of infant babbling

The first two months of infancy are characterized by a *phonation stage* (see Oller 1980, and Stark, 1980, for reviews of infant babbling), during which speech-like vocalizations are only rarely exhibited. The few speech-like sounds that can be observed consist largely of phonation with the mouth closed or nearly closed. The next developmental phase, occurring from 2 to 3 months of age, is known as the *"goo" stage* and is characterized by the production of crude syllable-like sequences composed mostly of velar consonant-like elements in combination with vowel-like elements. By 4 to 6 months old, most infants enter the *expansion stage*, characterized by the production of several new sound types, including labiolingual and bilabial trills, growls, and squeals. The expansion stage may also contain some of *marginal babbling,* consisting of vocal tract closures in combination with better-formed vowel-like utterances. Seven months of age sees most infants entering the *canonical or reduplicated babbling stage*, in which syllables with adult-like timing characteristics emerge. During this stage, many utterances consist of reduplicated syllables such as "bababa". The *nonreduplicated babbling* stage follows at around 10 months old; it is characterized by the use of different consonants and vowels within the same babbling sequence (e.g., "dadabi"). It has been suggested (MacNeilage and Davis 1990) that during the nonreduplicated babbling stage infants begin learning how to produce the phonemes of their native language.

An important feature of this developmental sequence is that many non-speech vocalizations and articulator movements occur well before the onset of frequent speech sounds. It is this observation that motivates the two learning stages of the DIVA model. In the first stage, sensory-motor relationships between the motor, somatosensory, and auditory systems are learned. In a sense, this stage consists of learning about the biophysics of the vocal tract; that is, the infant learns the sensory consequences of various oromotor actions. In the second stage, individual speech sounds from the native language are learned. While these stages are typically carried out sequentially in model simulations for convenience, the real speech motor learning process is not so discrete (e.g., de Boysson-Bardies, Sagart, and Durand 1984; de Boysson-Bardies et al. 1989; Mitchell and Kent 1990) and involves processes not addressed in computer simulations of DIVA. Table 1 provides an overview of these processes, which are detailed in the following paragraphs.

Development of auditory and somatosensory maps

The ability to produce the speech sounds of a language depends heavily on the ability to perceive these sounds. Auditory representations of speech signals (corresponding to the DIVA auditory state and auditory error maps) show signs of language specificity in infants as young as 6 months of age (e.g., Kuhl et al. 1992). This likely reflects modifications in auditory cortical neuronal responses to optimally capture the auditory signatures of the native language. This developmental process likely does not require knowledge of the phonological units that make up the language, as it occurs at a very early stage of development (see row 1 of Table 1). The shaping of auditory representations can instead be driven by the statistical nature of the acoustic signals experienced by the infant (e.g., Guenther and Gjaja 1996; Guenther et al. 1999).

The somatosensory representations of the speech network, corresponding to the somatosensory state map in Figure 1, must also undergo development. Unlike auditory signals for speech, the somatosensory patterns associated with the sounds of a language cannot be learned by listening to native speakers. Thus, development of the somatosensory maps for speech likely lags behind development of auditory maps during the very early stages of infancy, at a time when articulations are limited. Once the infant starts producing more speech-like articulatory movements in the expansion, canonical babbling, and nonreduplicated babbling stages, their somatosensory maps likely become increasingly sensitive to the somatosensory patterns proceeding from these movements (row 2 of Table 1).

Development of sensory-motor transformations

The first movements of speech-related body parts begin almost immediately after birth, when an infant uses their vocal folds and respiratory system to cry and their lips, jaw, and tongue to feed. These movements generate somatosensory feedback and often auditory feedback as well, providing opportunities for the infant's brain to learn about sensory consequences of oromotor actions. Our motor systems have the ability to anticipate sensory consequences of movements commanded by motor cortical activity. Tuning of these sensory-motor predictions, often referred to as forward models, likely begins with early non-speech actions, then accelerates as the infant creates more and more speech-like utterances as they move through the goo, expansion, canonical, and nonreduplicated babbling stages (rows 3, 4, and 5 in Table 1).

The articulatory movements which occur during infant babbling can also be used to tune transformations in the reverse direction, that is, sensory-to-motor transformations, or inverse models. These transformations consist of learned mappings between auditory and somatosensory representations of ongoing vocalizations and articulator movements that produce them. Prior to the development of auditory and somatosensory targets for speech sounds, nodes in the auditory and somatosensory error maps are not yet signaling "errors" *per se*; these nodes instead represent changes (velocities) in the auditory and somatosensory state that occur due to ongoing movements of speech articulators. This combination of motor activations and resulting sensory velocities enable the tuning of auditory-motor and

somato-motor transformations well before an infant develops awareness of phonological units such as phonemes and words.

Later, as auditory and somatosensory targets are learned, the nodes in the auditory and somatosensory error maps stop reflecting ongoing changes in the sensory state and begin to reflect desired sensory changes (i.e., sensory errors, which can be thought of as desired sensory velocities for reaching the target). This development, which can be inferred to have occurred when infants begin to produce language-specific speech sounds, is reflected in the DIVA model by the transition from the babbling phase to the imitation phase, though the model does not simulate specific mechanisms for the cause of this transition. Some continued tuning of sensory-motor transformations likely continues into adulthood; evidence for such plasticity is provided by adaptation to somatosensory feedback perturbations (e.g., Houde and Jordan 1998; Golfinopoulos et al. 2011; Lametti, Nasir, and Ostry 2012).

Speech recognition and phonological target acquisition

The learning processes described thus far do not require any knowledge of the distinct phonemes, syllables, or words of a language. Instead, they tune transformations between the largely continuous motor, somatosensory, and auditory spaces without regard for the discrete phonological units that make up a language. These transformations form the essential elements of the feedback control system schematized in Figure 1.

The ultimate goal of the speech motor system is, however, to produce these discrete speech sounds of the native language. Before a child can learn to articulate these sounds, it is required that they learn how to parse continuous auditory signals into discrete phonological categories such as words, syllables, and phonemes. This learning process corresponds to tuning of the speech recognition system and speech sound map in Figure 1. These learning processes (row 6 in Table 1) fall under the domain of speech perception and are not currently implemented in computer simulations of the DIVA model. Instead, speech sounds are presented to the model for learning; these sounds take the form of time-varying auditory signals (in particular, formant frequencies). Note that conscious awareness of phonemes is not a prerequisite for learning to produce phoneme strings; indeed, infants and children successfully learn words like “cat” and “hat” that differ only by a single phoneme despite not yet being consciously aware of phoneme units.

Development of sensory targets and feedforward control

As infants acquire auditory targets corresponding to phonemes and syllables, their brains store information about the sensory signals making up these objectives of speech motor output (row 7 in Table 1). The infant will then try to replicate these auditory targets. Projections to the auditory target map from the speech sound map encode these time-varying auditory targets for sounds represented in the speech sound map, so that these targets can be activated later during production of the corresponding sounds.

Infants have been reported to imitate caregivers’ vocalizations as early as 2 months old (Kuhl and Meltzoff 1996; Kokkinaki & Kugiumutzakis 2000; Gratier and Devouche 2011), while other accounts argue that this capacity emerges closer to 1 year of age (Jones 2009). These initial utterances enable the infant to learn feedforward commands for producing

these sounds on their own (row 8 in Table 1). Within the DIVA model, these feedforward commands are stored in synaptic projections from the speech sound map to the primary motor cortical areas, both directly and via a cortico-cerebellar loop.

Finally, after an infant can successfully produce speech sounds, the infant's brain develops a somatosensory target map containing representations of the somatic sensations created by accurately producing the sound (row 9 in Table 1). These targets are used by the somatosensory feedback control system to rapidly detect and correct production errors in ongoing utterances.

Computational modeling of developmental speech disorders

In addition to modeling normal development of speech production, variations of DIVA have also been used to simulate possible mechanisms of childhood disorders that affect speech production. Max et al. (2004) used mechanisms from DIVA to propose an account of developmental stuttering caused by dysfunctional use of auditory feedback. Subsequent simulation studies implemented this hypothesis (Civier, Tasko, and Guenther 2010), as well as alternative possible causes of the disorder (Civier et al. 2013). The neural etiology of childhood apraxia of speech has been addressed by DIVA modeling, in a study that simulated the disorder as resulting from impaired feedforward signaling (Terband et al. 2009; Miller and Guenther 2021). A recent application of the model used it to explore motor and auditory processing in children with autism spectrum disorder (Chenausky et al. 2021). A promising future direction for similar investigations may be the use of LaDIVA, a modification of the model which incorporates detailed laryngeal physiology, for understanding voice disorders such as pediatric dysphonia (Weerathunge et al. 2022).

III. Sequencing of speech motor programs

The previous sections discussed how the DIVA model simulates production of single speech motor programs and how these programs are learned and refined. Here we describe an extension to the DIVA model called the Gradient Order DIVA (GODIVA) model (Bohland, Bullock, and Guenther 2010) that describes the neural processes underlying the buffering and sequential production of longer utterances consisting of multiple speech sounds, such as phrases or sentences. In infancy, the capacity for rudimentary speech sound sequencing begins to manifest during nonreduplicated babbling (Levitt and Utman 1992; Nathani, Ertmer, and Stark 2006). GODIVA provides a description for developmental processes underlying the learning of these abilities. Before exploring these mechanisms, we give an overview of the components of the model.

Neural components of the GODIVA model

Figure 2 illustrates a simplified schematic of the GODIVA model. The model consists of two basal ganglia-thalamo-cortical loops (shaded regions in the figure): a *motor loop* (whose components are shared with the DIVA model) responsible for initiating and terminating speech motor programs, and a *planning loop* that forms a phonological working memory that buffers upcoming speech sounds. The planning loop involves the posterior inferior frontal sulcus (pIFS) in lateral prefrontal cortex and the presupplementary motor area

(preSMA) in the medial premotor cortex working in concert with the basal ganglia via projections to the head of the caudate nucleus, whereas the motor loop involves vPMC and SMA working in concert with the basal ganglia via projections to the putamen.

The model's cortical components can also be divided into medial and lateral cortical regions (indicated by dashed boxes in Figure 2), which represent distinct aspects of the speech utterance. One set of structures, the left lateral cortical areas pIFS and vPMC, contains representations of the speech sequence's phonological content (hypothesized to reside in left pIFS) and corresponding motor programs (hypothesized to reside in left vPMC). A second set, the medial premotor areas preSMA and SMA, are responsible for the metrical structure of the phonological sequence. Specifically, preSMA is hypothesized to contain a representation of syllabic frame structure and metrical patterning for an upcoming utterance, whereas SMA contains an initiation map (as in DIVA) that is responsible for turning on and turning off individual speech motor programs at particular instants in time. The planning loop regions preSMA and pIFS in GODIVA both use a gradient order working memory representation in which nodes representing actions to be produced sooner have higher activation levels than those to be produced later; such a representation has been proposed in prior computational models of working memory and sequencing (e.g., Lashley, 1951; Grossberg, 1978; Houghton, 1990; Houghton and Hartley, 1996). The following subsections provide further detail regarding the model's medial and lateral streams.

Processing of sequential structure in medial premotor cortex

The GODIVA model posits that preSMA contains a representation of the global metrical structure of an upcoming speech utterance, whereas SMA is primarily responsible for initiating the motor execution of speech articulations. The SMA and preSMA elements in GODIVA are inspired in part by single unit electrophysiological studies of action sequencing in non-human primates. For example, Shima and Tanji (2000) trained macaque monkeys to perform different sequences of three hand/arm movements (e.g., push-pull-turn) while recording from neurons in SMA and preSMA. Broadly speaking, neurons in SMA were more closely tied to particular movements, whereas neurons in preSMA often represented more global aspects of the full sequence, for example neurons that fired at the beginning of only one particular three-movement sequence, or neurons that fired during production of the second (or first, or third) movement of the sequence regardless of whether the movement was a push, pull, or turn. Subsequent human neuroimaging studies found a corresponding association between speech sequence complexity and preSMA activation (Bohland and Guenther 2006; Rong et al. 2018).

In GODIVA, preSMA nodes represent the syllable frame structure and stress patterning of the utterance, which determine the utterance's metrical structure. Projections from preSMA nodes to SMA are responsible (in concert with the basal ganglia, as described below) for activating and deactivating the proper SMA initiation map nodes (each of which launches a distinct motor program) in the proper order and with the proper stress. In this way, the medial stream of the GODIVA model dictates the metrical structure/tempo of a multi-sound utterance.

Phonological content buffering in lateral prefrontal cortex

According to GODIVA, pIFS contains a phonological content buffer for temporarily storing the phonological units of an upcoming utterance. This function is assigned to left IFS based on demonstrations of its role in working memory (Kerns et al. 2004; Gabrieli, Poldrack, and Desmond 1998; Kumar et al. 2016), particularly verbal working memory (Rottschy et al. 2012), as well as its encoding of phonological identity and complexity (Poldrack et al. 1999; Bohland and Guenther 2006; Myers et al. 2009). Activity in this region also is associated with acquisition of phonetic categories in infants during the first year of life (Imada et al. 2006).

Each node in the phonological content buffer represents a different phonological unit (e.g., a phoneme or consonant cluster). The order of upcoming speech sounds to be produced is represented by the gradient of activity across these nodes. GODIVA, like the DIVA model, implements speech sound map nodes residing in vPMC. Once pIFS selects the next motor program to execute, as determined by the highest-activity node in its phonological buffer, this selection is transmitted to left vPMC via projections from pIFS. Execution of the motor program begins at the instant the corresponding SMA initiation map node is activated (at which time the sound's representation is deleted from the pIFS phonological content buffer), and the motor program terminates when the initiation map node activity is extinguished.

Motor sequence chunking and automatization in the basal ganglia loop

We propose that, early in development, the working memory areas preSMA and pIFS must be heavily involved in the speech sequencing process since frequently occurring sequences haven't yet been "automated" by transferring control of the sequence to subcortical structures. In GODIVA, if a particular movement sequence is repeated many times, nodes in the basal ganglia learn to recognize the sensorimotor context for initiating the individual items in the sequence. After learning, the sequence is represented by its own speech sound map node, and activating this node leads to readout of the learned movement sequence. The learning process is schematized in Figure 3.

The cortico-basal ganglia motor loop accomplishes this automation of frequently used speech sequences in early childhood by encoding these sequences as "chunks" with their own optimized motor programs. This chunking would reduce the processing load on prefrontal and premotor cortical areas (Alm 2004; Redgrave et al. 2010). For example, the speech motor system of a young child might attempt to produce the word "snow" (Figure 3, Panel A). vMC contains nodes encoding articulatory gestures (labeled G) for the phonemes /s/, /n/, and /ō/. Each phonemic gesture has a corresponding cell in the SMA initiation map (labeled I) that is responsible for initiating the gesture via projections to vMC. During this early stage of development, vPMC does not contain a motor program for the entire syllable /snō/. Instead, the syllable is represented by individual motor programs for each phoneme that must be activated independently via inputs from the IFS phonological buffer. Similarly, preSMA and pIFS contain only phonemic elements, not larger units such as consonant clusters.

At this stage, production of the word requires activation of the nodes /s/, /n/, and /ō/ in the phonological content buffer in pIFS, as well as the structural representation for /snō/ in the sequential structure buffer in preSMA. Projections from pIFS sequentially activate the vPMC nodes corresponding to the motor programs for /s/, /n/, and /ō/. Projections from these vPMC nodes sequentially activate the matching gestural nodes in vMC. The timing of this sequential activation process is determined by the medial premotor areas. PreSMA-to-SMA projections activate nodes in the initiation map for the individual phonemes in the proper order and with the proper timing. Once a motor program has been completed, the pIFS, vPMC, and pIFS nodes for that program's elements are deactivated, allowing the next motor program to commence.

Panel B of Figure 3 schematizes the production of /snō/ at a more mature stage of development. At this stage, vPMC contains a motor program for the entire syllable /snō/, with subcortical loops through the cerebellum (green dashed arrows) effectively taking over coordination of the individual motor gestures. The importance of the cerebellum for vocal sequence learning has been empirically supported by pediatric clinical studies and animal lesion models (Ziegler and Ackermann 2017; Pidoux et al. 2018; Glickstein 1994). Once these cortical-subcortical loops are established, working memory buffers in preSMA and pIFS will contain cluster-sized sub-syllabic units, thereby reducing the number of items that have to be stored in working memory for /snō/. The task of initiating the gesture for /n/ in /snō/ now gets carried out by the basal ganglia motor loop (red dashed arrow) instead of preSMA.

This learning process reduces the number of pIFS, preSMA, and vPMC nodes that must be activated to produce the word. The required number of cortico-cortical connections (black arrows) has decreased substantially, having been replaced by subcortical communications through the cerebellum (green arrows) and basal ganglia (red arrows). Evidence for speech learning-related reductions in processing load has been demonstrated by neuroimaging studies of nonnative consonant cluster learning (Segawa et al. 2015; Masapollo et al. 2021).

IV. Summary

This review described neuro-computational approaches for modeling infant and child speech motor development. We first provided an overview of the DIVA model, which characterizes feedforward and feedback mechanisms of speech production controlled by a network of cortical and subcortical loops. The feedforward control system is thought to involve cortico-cortical projections from premotor to motor cortex, as well as contributions from the cerebellum. The auditory and somatosensory feedback control systems monitor the perceptual consequences of speech output, which are compared to sensory predictions transmitted from premotor cortex to higher-order sensory areas. These sensory areas compute error signals, which are sent to motor cortex as corrective motor commands.

We described how early stages of speech motor learning can be simulated with the DIVA model. Speech motor development involves a number of learning processes occurring in a quasi-parallel fashion. Infant babbling and other vocalizations begin tuning forward maps which map motor outputs to resulting auditory and somatosensory perceptions. Auditory

maps develop in a way that highlights important acoustic distinctions in a language and de-emphasizes irrelevant distinctions. Analogously, somatosensory maps become sensitive to the tactile and proprioceptive feedback patterns that occur when producing sounds from the native language. Auditory targets for speech sound "chunks" such as phonemes, syllables, and words are formed by monitoring the environment for native language samples, and feedforward commands are tuned as a child attempts to produce these sound chunks.

Next, we addressed computational modeling of a more advanced stage of child speech development, in which longer phonological sequences such as phrases or sentences are produced. Modeling of these processes uses the Gradient Order DIVA (GODIVA) model. High-level language processing regions maintain temporary stores of upcoming phonological content and metrical structure in competitive queues. These regions control the output of the downstream initiation maps and speech sound maps to produce sequences of speech sounds. GODIVA also describes a mechanism of speech sequence learning, or chunking, via cortico-basal ganglia loops. Frequently produced motor sequences that formerly required cortical control for every sequential step are automated into syllabic motor programs controlled mostly by the basal ganglia and cerebellum, reducing cortical processing load as the child proceeds through speech development.

Acknowledgements

This research was funded by the following grants from the National Institutes of Health: R01 DC007683 (F. Guenther, PI), R01 DC016270 (C. Stepp and F. Guenther, MPIs), U01 NS117836, (M. Richardson, PI), R01 019354 (M. Long, PI).

References

Alm Per A. 2004. "Stuttering and the Basal Ganglia Circuits: A Critical Review of Possible Relations." Journal of Communication Disorders 37 (4): 325–69. [PubMed: 15159193]

Anthony Jason L, and Francis David J. 2005. "Development of Phonological Awareness." Current Directions in Psychological Science 14 (5): 255–59.

Anthony JL, Lonigan CJ, Driscoll K, Phillips BM and Burgess SR 2003. "Phonological sensitivity: A quasi-parallel progression of word structure units and cognitive operations." Reading Research Quarterly 38(4): 470–487.

Benedict Helen. 1979. "Early Lexical Development: Comprehension and Production." Journal of Child Language 6 (2): 183–200. [PubMed: 468932]

Bohland Jason W., Bullock Daniel, and Guenther Frank H.. 2010. "Neural Representations and Mechanisms for the Performance of Simple Speech Sequences." Journal of Cognitive Neuroscience 22 (7): 1504–29. 10.1162/jocn.2009.21306. [PubMed: 19583476]

Bohland Jason W., and Guenther Frank H.. 2006. "An FMRI Investigation of Syllable Sequence Production." NeuroImage 32 (2): 821–41. 10.1016/j.neuroimage.2006.04.173. [PubMed: 16730195]

Bénédicte de Boysson-Bardies, Hallé Pierre, Sagart Laurent, and Durand Catherine. 1989. "A Crosslinguistic Investigation of Vowel Formants in Babbling." Journal of Child Language 16 (1): 1–17. [PubMed: 2925806]

Bénédicte de Boysson-Bardies, Sagart Laurent, and Durand Catherine. 1984. "Discernible Differences in the Babbling of Infants According to Target Language." Journal of Child Language 11 (1): 1–15. [PubMed: 6699104]

Callan DE, Kent RD, Guenther FH, and Vorperian HK. 2000. "An Auditory-Feedback-Based Neural Network Model of Speech Production That Is Robust to Developmental Changes in the Size and Shape of the Articulatory System." Journal of Speech, Language, and Hearing Research: JSLHR 43 (3): 721–36. 10.1044/jslhr.4303.721. [PubMed: 10877441]

Chenausky Karen V., Brignell Amanda, Morgan Angela T., Norton Andrea C., Tager-Flusberg Helen B., Schlaug Gottfried, and Guenther Frank H.. 2021. “A Modeling-Guided Case Study of Disordered Speech in Minimally Verbal Children With Autism Spectrum Disorder.” American Journal of Speech-Language Pathology 30 (3 Suppl): 1542–57. 10.1044/2021_AJSLP-20-00121. [PubMed: 33852328]

Civier Oren, Bullock Daniel, Max Ludo, and Guenther Frank H.. 2013. “Computational Modeling of Stuttering Caused by Impairments in a Basal Ganglia Thalamo-Cortical Circuit Involved in Syllable Selection and Initiation.” Brain and Language 126 (3): 263–78. 10.1016/j.bandl.2013.05.016. [PubMed: 23872286]

Civier Oren, Tasko Stephen M., and Guenther Frank H.. 2010. “Overreliance on Auditory Feedback May Lead to Sound/Syllable Repetitions: Simulations of Stuttering and Fluency-Inducing Conditions with a Neural Model of Speech Production.” Journal of Fluency Disorders 35 (3): 246–79. 10.1016/j.jfludis.2010.05.002. [PubMed: 20831971]

DeFelipe J and Fariñas I, 1992. The pyramidal neuron of the cerebral cortex: morphological and chemical characteristics of the synaptic inputs. Progress in neurobiology, 39(6), pp.563–607. [PubMed: 1410442]

Doyon J, Bellec P, Amsel R, Penhune V, Monchi O, Carrier J, Lehéricy S and Benali H, 2009. Contributions of the basal ganglia and functionally related brain structures to motor learning. Behavioural brain research, 199(1), pp.61–75. [PubMed: 19061920]

Dubois J, Dehaene-Lambertz G, Kulikova S, Poupon C, Hüppi PS, & Hertz-Pannier L (2014). The early development of brain white matter: a review of imaging studies in fetuses, newborns and infants. Neuroscience, 276, 48–71. [PubMed: 24378955]

Fromkin Victoria A. 1971. “The Non-Anomalous Nature of Anomalous Utterances.” Language 47 (1): 27–52. 10.2307/412187.

Gabrieli John D. E., Poldrack Russell A., and Desmond John E.. 1998. “The Role of Left Prefrontal Cortex in Language and Memory.” Proceedings of the National Academy of Sciences 95 (3): 906–13. 10.1073/pnas.95.3.906.

Gervain Judit, and Mehler Jacques. 2010. “Speech Perception and Language Acquisition in the First Year of Life.” Annual Review of Psychology 61: 191–218.

Glickstein Mitchell. 1994. “Cerebellar Agenesis.” Brain 117 (5): 1209–12. 10.1093/brain/117.5.1209. [PubMed: 7953600]

Golfinopoulos E, Tourville JA, and Guenther FH. 2010. “The Integration of Large-Scale Neural Network Modeling and Functional Brain Imaging in Speech Motor Control.” NeuroImage, Computational Models of the Brain, 52 (3): 862–74. 10.1016/j.neuroimage.2009.10.023.

Golfinopoulos Elisa, Tourville Jason A, Bohland Jason W, Ghosh Satrajit S, Nieto-Castanon Alfonso, and Guenther Frank H. 2011. “FMRI Investigation of Unexpected Somatosensory Feedback Perturbation during Speech.” Neuroimage 55 (3): 1324–38. [PubMed: 21195191]

Gout Ariel, Christophe Anne, and Morgan James L. 2004. “Phonological Phrase Boundaries Constrain Lexical Access II. Infant Data.” Journal of Memory and Language 51 (4): 548–67.

Gratier M and Devouche E 2011. “Imitation and repetition of prosodic contour in vocal interaction at 3 months.” Developmental Psychology 47(1), 67. [PubMed: 21244150]

Graven Stanley N, and Browne Joy V. 2008. “Auditory Development in the Fetus and Infant.” Newborn and Infant Nursing Reviews 8 (4): 187–93.

Grossberg S (1978a). A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans. Prog Theor Biol, 5, 233–374.

Guenther Frank H. 1995. “Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production.” Psychological Review 102 (3): 594. [PubMed: 7624456]

Guenther Frank H. 2016. Neural Control of Speech. Mit Press.

Guenther Frank H, and Gjaja Marin N. 1996. “The Perceptual Magnet Effect as an Emergent Property of Neural Map Formation.” The Journal of the Acoustical Society of America 100 (2): 1111–21. [PubMed: 8759964]

Guenther Frank H, Husain Fatima T, Cohen Michael A, and Shinn-Cunningham Barbara G. 1999. “Effects of Categorization and Discrimination Training on Auditory Perceptual Space.” The Journal of the Acoustical Society of America 106 (5): 2900–2912. [PubMed: 10573904]

Hao YC, & de Jong K (2016). Imitation of second language sounds in relation to L2 perception and production. Journal of Phonetics, 54, 151–168.

Hashimoto Y and Sakai KL, 2003. Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. Human brain mapping, 20(1), pp.22–28. [PubMed: 12953303]

Hickok G (2014). The architecture of speech production and the role of the phoneme in speech processing. Language, Cognition and Neuroscience, 29(1), 2–20.

Hikosaka O, Nakamura K, Sakai K and Nakahara H, 2002. Central mechanisms of motor skill learning. Current opinion in neurobiology, 12(2), pp.217–222. [PubMed: 12015240]

Houde John F, and Jordan Michael I. 1998. “Sensorimotor Adaptation in Speech Production.” Science 279 (5354): 1213–16. [PubMed: 9469813]

Houghton G (1990). The problem of serial order: a neural network model of sequence learning and recall. In Dale R, Mellish C, and Zock M (Eds.), Current research in natural language generation (pp. 287–319). San Diego: Academic Press.

Houghton George, and Hartley Tom. 1996. “Parallel Models of Serial Behaviour: Lashley Revisited.” Psyche: An Interdisciplinary Journal of Research on Consciousness 2: No Pagination Specified-No Pagination Specified.

Imada Toshiaki, Zhang Yang, Cheour Marie, Taulu Samu, Ahonen Antti, and Kuhl Patricia K.. 2006. “Infant Speech Perception Activates Broca’s Area: A Developmental Magnetoencephalography Study.” NeuroReport 17 (10): 957–62. 10.1097/01.wnr.0000223387.51704.89. [PubMed: 16791084]

Jones SS, 2009. “The development of imitation in infancy.” Philosophical Transactions of the Royal Society B: Biological Sciences 364(1528): 2325–2335.

Jusczyk Peter W, Cutler Anne, and Redanz Nancy J. 1993. “Infants’ Preference for the Predominant Stress Patterns of English Words.” Child Development 64 (3): 675–87. [PubMed: 8339688]

Jusczyk Peter W, and Hohne Elizabeth A. 1997. “Infants’ Memory for Spoken Words.” Science 277 (5334): 1984–86. [PubMed: 9302291]

Kehoe Margaret, and Stoel-Gammon Carol. 1997. “The Acquisition of Prosodic Structure: An Investigation of Current Accounts of Children’s Prosodic Development.” Language 73 (1): 113–44. 10.2307/416597.

Kerns John G, Cohen Jonathan D, Stenger V. Andrew, and Carter Cameron S. 2004. “Prefrontal Cortex Guides Context-Appropriate Responding during Language Production.” Neuron 43 (2): 283–91. 10.1016/j.neuron.2004.06.032. [PubMed: 15260963]

Kessler Brett, and Treiman Rebecca. 1997. “Syllable Structure and the Distribution of Phonemes in English Syllables.” Journal of Memory and Language 37 (3): 295–311. 10.1006/jmla.1997.2522.

Kokkinaki T and Kugiumutzakis G, 2000. “Basic aspects of vocal imitation in infant-parent interaction during the first 6 months.” Journal of reproductive and infant psychology 18(3):173–187.

Kostović I, & Jovanov-Milošević N (2006, December). The development of cerebral connections during the first 20–45 weeks’ gestation. In seminars in fetal and neonatal medicine (Vol. 11, No. 6, pp. 415–422). WB Saunders. [PubMed: 16962836]

Kozhevnikov VA, and Chistovich LA. 1965. Speech: Articulation and Perception. Speech: Articulation and Perception. Oxford, England: Nauka.

Kuhl PK and Meltzoff AN, 1996. “Infant vocalizations in response to speech: Vocal imitation and developmental change.” The journal of the Acoustical Society of America 100(4): 2425–2438. [PubMed: 8865648]

Kuhl Patricia K, Williams Karen A, Lacerda Francisco, Stevens Kenneth N, and Lindblom Björn. 1992. “Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age.” Science 255 (5044): 606–8. [PubMed: 1736364]

Kumar Sukhbinder, Joseph Sabine, Gander Phillip E., Barascud Nicolas, Halpern Andrea R., and Griffiths Timothy D.. 2016. “A Brain System for Auditory Working Memory.” Journal of Neuroscience 36 (16): 4492–4505. 10.1523/JNEUROSCI.4341-14.2016. [PubMed: 27098693]

Lametti Daniel R., Nasir Sazzad M., and Ostry David J.. 2012. “Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback.”

Journal of Neuroscience 32 (27): 9351–58. 10.1523/JNEUROSCI.0404-12.2012. [PubMed: 22764242]

Lashley KS (1951). The problem of serial order in behavior. In: Jeffress L (Ed.), Cerebral mechanisms in behavior. New York: Wiley, pp 112–136.

Li LY, Ji XY, Liang F, Li YT, Xiao Z, Tao HW and Zhang LI, 2014. A feedforward inhibitory circuit mediates lateral refinement of sensory representation in upper layer 2/3 of mouse primary auditory cortex. Journal of Neuroscience, 34(41), pp.13670–13683. [PubMed: 25297094]

Levelt Willem J. M. 1993. Speaking: From Intention to Articulation. 10.7551/mitpress/6393.001.0001.

Levitt Andrea G., and Aydelott Utman Jennifer G.. 1992. “From Babbling towards the Sound Systems of English and French: A Longitudinal Two-Case Study*.” Journal of Child Language 19 (1): 19–49. 10.1017/S0305000900013611. [PubMed: 1551932]

Maclean M, Bryant P and Bradley L 1987. “Rhymes, nursery rhymes, and reading in early childhood.” Merrill-Palmer Quarterly (1982-): 255–281.

MacNeilage Peter F, and Davis Barbara L. 1990. “Acquisition of Speech Production: The Achievement of Segmental Independence.” In Speech Production and Speech Modelling, 55–68. Springer.

Masapollo Matthew, Segawa Jennifer A., Beal Deryk S., Tourville Jason A., Alfonso Nieto-Castañón Matthias Heyne, Frankford Saul A., and Guenther Frank H.. 2021. “Behavioral and Neural Correlates of Speech Motor Sequence Learning in Stuttering and Neurotypical Speakers: An FMRI Investigation.” Neurobiology of Language 2 (1): 106–37. 10.1162/nol_a_00027. [PubMed: 34296194]

Mattys Sven L, and Jusczyk Peter W. 2001. “Do Infants Segment Words or Recurring Contiguous Patterns?” Journal of Experimental Psychology: Human Perception and Performance 27 (3): 644. [PubMed: 11424651]

Mattys Sven L., Jusczyk Peter W., Luce Paul A., and Morgan James L.. 1999. “Phonotactic and Prosodic Effects on Word Segmentation in Infants.” Cognitive Psychology 38 (4): 465–94. 10.1006/cogp.1999.0721. [PubMed: 10334878]

Max Ludo, Guenther Frank H., Gracco Vincent L., Ghosh Satrajt S., and Wallace Marie E.. 2004. “Unstable or Insufficiently Activated Internal Models and Feedback-Biased Motor Control as Sources of Dysfluency: A Theoretical Model of Stuttering.” Contemporary Issues in Communication Science and Disorders 31 (Spring): 105–22. 10.1044/cicsd_31_S_105.

Miller Hilary E., and Guenther Frank H.. 2021. “Modelling Speech Motor Programming and Apraxia of Speech in the DIVA/GODIVA Neurocomputational Framework.” Aphasiology 35 (4): 424–41. 10.1080/02687038.2020.1765307. [PubMed: 34108793]

Mitchell Pamela R., and Kent Raymond D.. 1990. “Phonetic Variation in Multisyllable Babbling*.” Journal of Child Language 17 (2): 247–65. 10.1017/S0305000900013751. [PubMed: 2380268]

Myers Emily B., Blumstein Sheila E., Walsh Edward, and Eliassen James. 2009. “Inferior Frontal Regions Underlie the Perception of Phonetic Category Invariance.” Psychological Science 20 (7): 895–903. 10.1111/j.1467-9280.2009.02380.x. [PubMed: 19515116]

Naskar S, Qi J, Pereira F, Gerfen CR and Lee S, 2021. Cell-type-specific recruitment of GABAergic interneurons in the primary somatosensory cortex by long-range inputs. Cell reports, 34(8), p.108774. [PubMed: 33626343]

Nathani Suneeti, Ertmer David J., and Stark Rachel E.. 2006. “Assessing Vocal Development in Infants and Toddlers.” Clinical Linguistics & Phonetics 20 (5): 351–69. 10.1080/02699200500211451. [PubMed: 16728333]

Nozari N, Dell GS, & Schwartz MF (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. Cognitive psychology, 63(1), 1–33. [PubMed: 21652015]

Okada K, Matchin W and Hickok G, 2018. Neural evidence for predictive coding in auditory cortex during speech production. Psychonomic bulletin & review, 25, pp.423–430. [PubMed: 28397076]

Oller DK 1980. “The Emergence of the Speech Capacity in Infancy.” In Child Phonology, edited by Yeni-komshian GRACEH, Kavanagh JAMESF, and Ferguson CHARLESA, 93–112. Academic Press. 10.1016/B978-0-12-770601-6.50011-5.

Ozker M, Doyle W, Devinsky O and Flinker A, 2022. A cortical network processes auditory error signals during human speech production to maintain fluency. PLoS Biology, 20(2), p.e3001493. [PubMed: 35113857]

Pidoux Ludivine, Pascale Le Blanc Carole Levenes, and Leblois Arthur. 2018. “A Subcortical Circuit Linking the Cerebellum to the Basal Ganglia Engaged in Vocal Learning.” Edited by Raymond Jennifer L and King Andrew J. ELife 7 (July): e32167. 10.7554/eLife.32167. [PubMed: 30044222]

Poldrack Russell A., Wagner Anthony D., Prull Matthew W., Desmond John E., Glover Gary H., and Gabrieli John D. E.. 1999. “Functional Specialization for Semantic and Phonological Processing in the Left Inferior Prefrontal Cortex.” NeuroImage 10 (1): 15–35. 10.1006/nimg.1999.0441. [PubMed: 10385578]

Redgrave Peter, Rodriguez Manuel, Smith Yoland, Rodriguez-Oroz Maria C., Lehericy Stephane, Bergman Hagai, Agid Yves, DeLong Mahlon R., and Obeso Jose A.. 2010. “Goal-Directed and Habitual Control in the Basal Ganglia: Implications for Parkinson’s Disease.” Nature Reviews Neuroscience 11 (11): 760–72. 10.1038/nrn2915. [PubMed: 20944662]

Robles S. Gil, Gatignol P, Capelle L, Mitchell M-C, and Duffau. 2005. “The Role of Dominant Striatum in Language: A Study Using Intraoperative Electrical Stimulations.” Journal of Neurology, Neurosurgery & Psychiatry 76 (7): 940–46. 10.1136/jnnp.2004.045948. [PubMed: 15965199]

Rong F, Isenberg AL, Sun E and Hickok G, 2018. The neuroanatomy of speech sequencing at the syllable level. PloS one, 13(10), p.e0196381. [PubMed: 30300341]

Rottschy C, Langner R, Dogan I, Reetz K, Laird AR, Schulz JB, Fox PT, and Eickhoff SB. 2012. “Modelling Neural Correlates of Working Memory: A Coordinate-Based Meta-Analysis.” NeuroImage 60 (1): 830–46. 10.1016/j.neuroimage.2011.11.050. [PubMed: 22178808]

Segawa Jennifer A., Tourville Jason A., Beal Deryk S., and Guenther Frank H.. 2015. “The Neural Correlates of Speech Motor Sequence Learning.” Journal of Cognitive Neuroscience 27 (4): 819–31. 10.1162/jocn_a_00737. [PubMed: 25313656]

Seidl Amanda, and Johnson Elizabeth K. 2006. “Infant Word Segmentation Revisited: Edge Alignment Facilitates Target Extraction.” Developmental Science 9 (6): 565–73. [PubMed: 17059453]

Shima Keisetsu, and Tanji Jun. 2000. “Neuronal Activity in the Supplementary and Presupplementary Motor Areas for Temporal Organization of Multiple Movements.” Journal of Neurophysiology 84 (4): 2148–60. 10.1152/jn.2000.84.4.2148. [PubMed: 11024102]

Stark Rachel E. 1980. “Stages of Speech Development in the First Year of Life.” In Child Phonology, 73–92. Elsevier.

Sun Yue, and Poeppel David. 2022. “Syllables and Their Beginnings Have a Special Role in the Mental Lexicon.” PsyArXiv. 10.31234/osf.io/c9tx2.

Terband H, Maassen B, Guenther FH, and Brumberg J. 2009. “Computational Neural Modeling of Speech Motor Control in Childhood Apraxia of Speech (CAS).” Journal of Speech, Language, and Hearing Research : JSLHR 52 (6): 1595–1609. 10.1044/1092-4388(2009/07-0283). [PubMed: 19951927]

Tourville Jason A., and Guenther Frank H.. 2011. “The DIVA Model: A Neural Theory of Speech Acquisition and Production.” Language and Cognitive Processes 26 (7): 952–81. 10.1080/01690960903498424. [PubMed: 23667281]

Treiman R, Fowler CA, Gross J, Berch D, and Weatherston S. 1995. “Syllable Structure or Word Structure? Evidence for Onset and Rime Units with Disyllabic and Trisyllabic Stimuli.” Journal of Memory and Language 34 (1): 132–55. 10.1006/jmla.1995.1007.

Treiman R and Zukowski A 1996. “Children’s sensitivity to syllables, onsets, rimes, and phonemes.” Journal of experimental child psychology 61(3): 193–215. [PubMed: 8636664]

Urrutia-Piñones J, Morales-Moraga C, Sanguinetti-González N, Escobar AP and Chiu CQ, 2022. Long-range gabaergic projections of cortical origin in brain function. Frontiers in Systems Neuroscience, 16.

Weerathunge Hasini R., Alzamendi Gabriel A., Cler Gabriel J., Guenther Frank H., Stepp Cara E., and Zañartu Matías. 2022. “LaDIVA: A Neurocomputational Model Providing Laryngeal Motor Control for Speech Acquisition and Production.” PLOS Computational Biology 18 (6): e1010159. 10.1371/journal.pcbi.1010159. [PubMed: 35737706]

Ziegler W, and Ackermann H. 2017. “Subcortical Contributions to Motor Speech: Phylogenetic, Developmental, Clinical.” Trends in Neurosciences 40 (8): 458–68. 10.1016/j.tins.2017.06.005. [PubMed: 28712469]

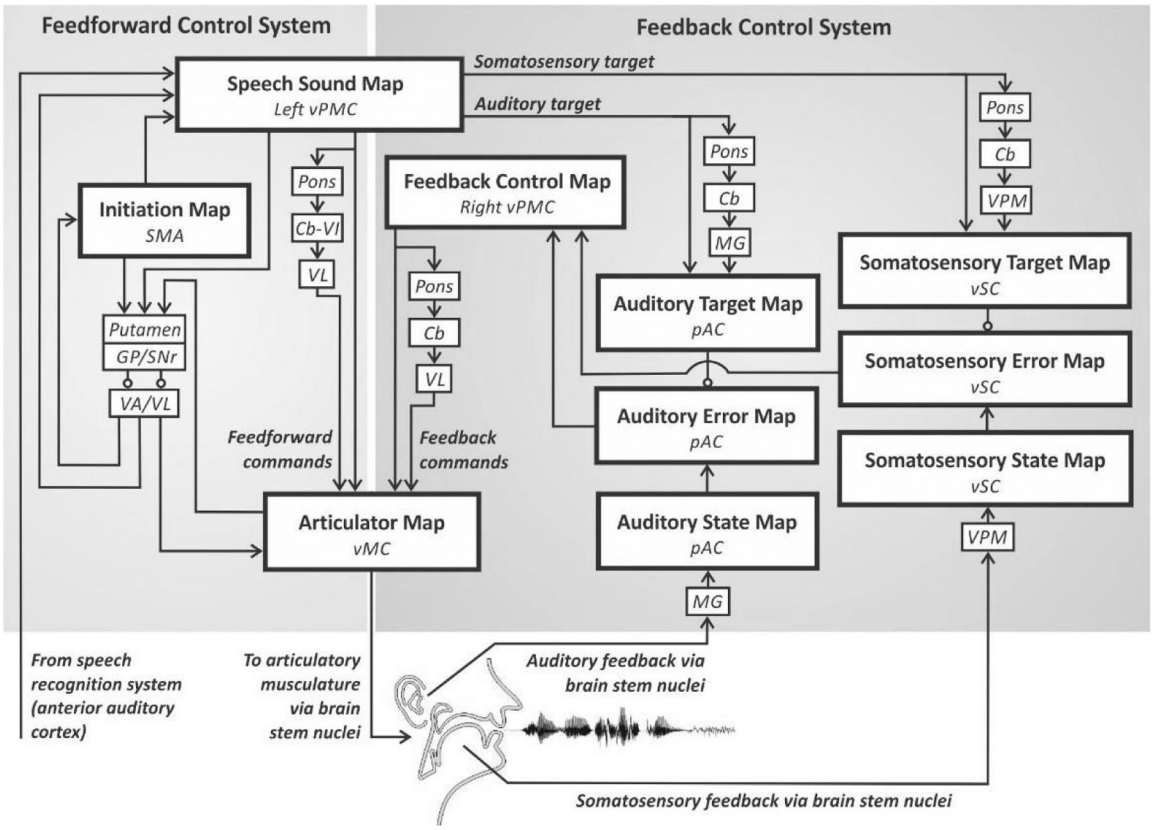

Figure 1.

Neural correlates of the DIVA model. The main neural output of the model is provided by the vMC Articulator Map, which integrates feedforward commands from VL and the Speech Sound Map with feedback commands from VL and the Feedback Control Map. [Abbreviations: Cb=cerebellum (specific lobule unknown); Cb-VI=cerebellum lobule VI; GP=globus pallidus; MG=medial geniculate nucleus of the thalamus; pAC=posterior auditory cortex; SMA=supplementary motor area; SNr=substantia nigra pars reticula; VA=ventral anterior nucleus of the thalamus; VL=ventral lateral nucleus of the thalamus; vMC=ventral motor cortex; VPM=ventral posterior medial nucleus of the thalamus; vPMC=ventral premotor cortex; vSC=ventral somatosensory cortex.]

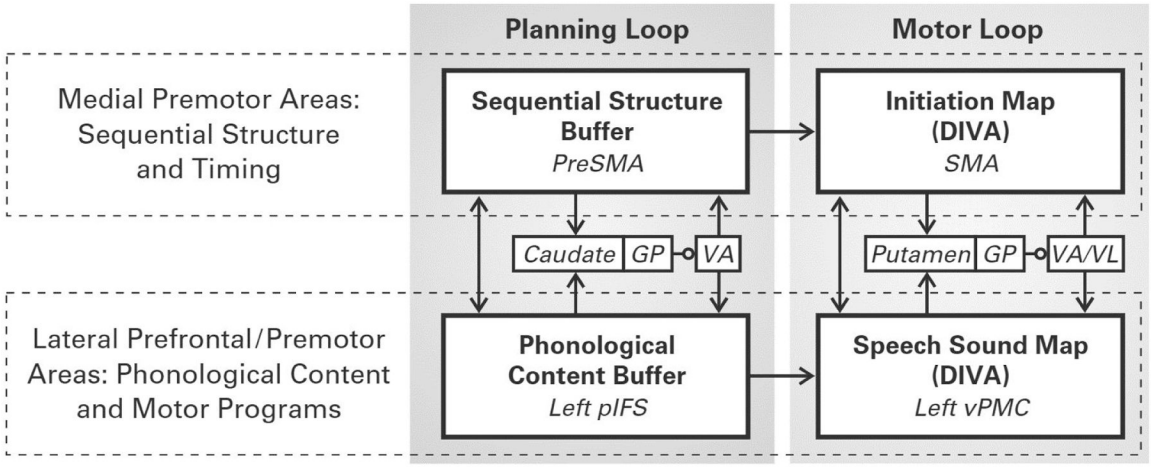

Figure 2.

Simplified schematic of the GODIVA network model for speech sequence production. [Abbreviations: GP, globus pallidus; pIFS, posterior inferior frontal sulcus; preSMA, presupplementary motor area; SMA, supplementary motor area; VA, ventral anterior thalamic nucleus; VL, ventral lateral thalamic nucleus; vPMC, ventral premotor cortex]

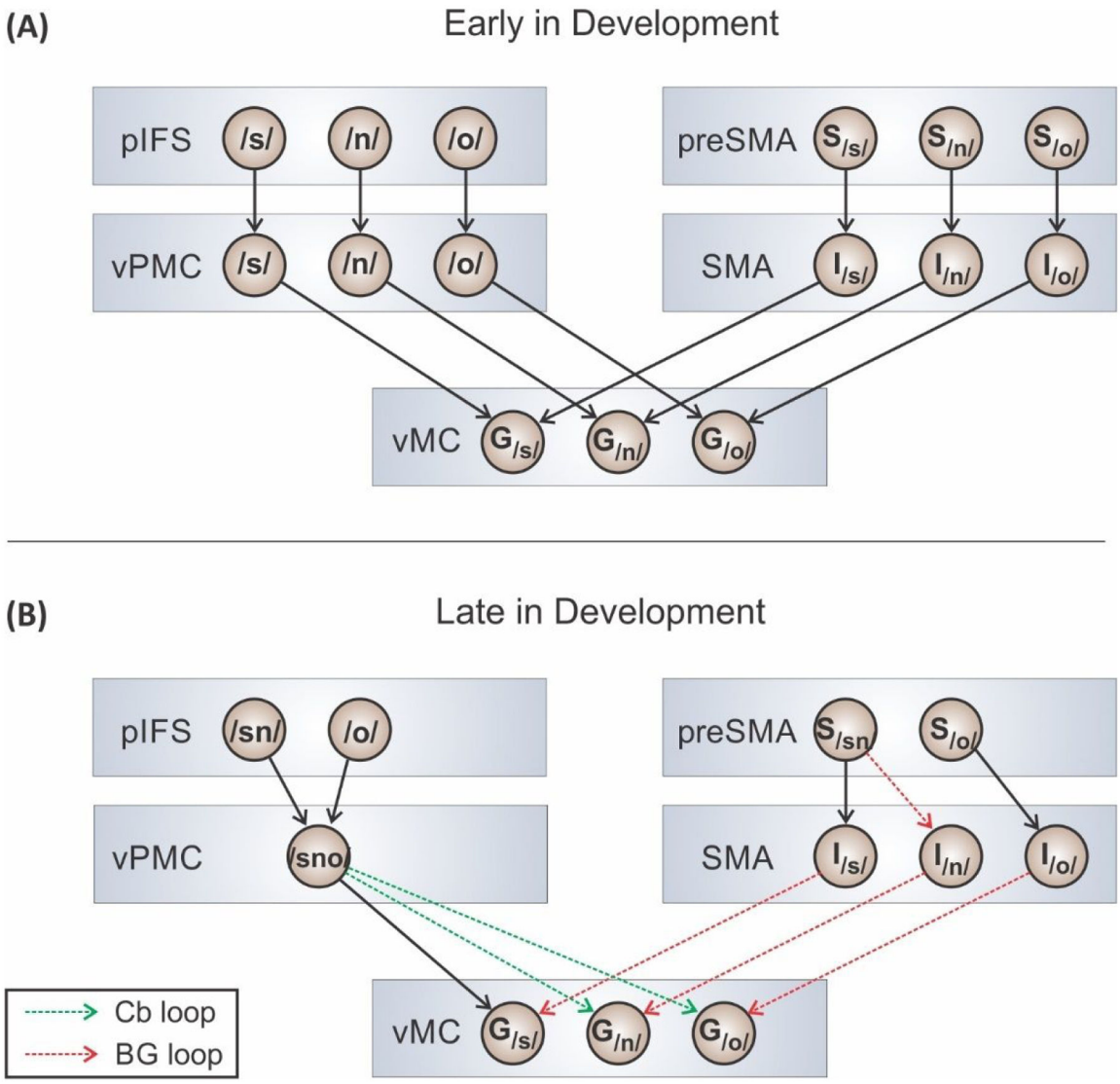

Figure 3.

Illustration of speech sequence learning via "chunking" in the GODIVA model. (A) Network involved in producing the word "snow" early in speech motor development. Cortico-cortical projections are indicated by black arrows. (B) Network involved in producing the word "snow" later in development. The development of basal ganglia (red dashed arrows) and cerebellar (green dashed arrows) loops allow for the use of fewer cortical nodes and projections. [Abbreviations: BG, basal ganglia; Cb, cerebellum; G, gestural node; I, initiation map node; pIFS, posterior inferior frontal sulcus; preSMA, presupplementary motor area; S, syllabic structure node; SMA, supplementary motor area; vMC, ventral primary motor cortex; vPMC, ventral premotor cortex]

Table 1.

Time-courses for development of the major capacities of the speech motor system. The estimated amount of learning occurring in a neural system within a given time window is indicated as being *Low*, *Medium*, or *High*.

Neural System	Age and Development Stage								
	0–1 mo. phonation	**2–3 mo. goo**	**4–6 mo. expansion**	**7–10 mo. canonical**	**10–12 mo. Non-reduplicated**	**1–2 yr. words**	**3–5 yr. sentences**	**6–18 yr. school**	**> 18 yr. mature**
1. Aud. state and error maps	Low	Low	Med	High	High	High	High	Med	Low
2. Som. state and error maps	Low	Low	Med	High	High	High	High	Med	Low
3. Aud.-motor transformations	Low	Low	Med	High	High	High	High	Med	Low
4. Som.-motor transformations	Low	Low	Med	High	High	High	High	Med	Low
5. Som.-aud. transformations	Low	Low	Med	High	High	High	High	Med	Low
6. Speech sound map			Low	Med	High	High	High	Med	Low
7. Aud. Target map			Low	Med	High	High	High	Med	Low
8. Feedforward commands			Low	Med	High	High	High	Med	Low
9. Som. target map			Low	Low	Med	High	High	Med	Low

Abbreviations: Aud.=auditory; Som.=somatosensory.