The DIVA model: A neural theory of speech acquisition and production

Jason A. Tourville
Department of Cognitive and Neural Systems, Boston University, Boston, MA, USA

Frank H. Guenther
Department of Cognitive and Neural Systems, Boston University, Boston, MA; Division of Health Sciences and Technology, Harvard University – Massachusetts Institute of Technology; Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, MA, USA

The DIVA model of speech production provides a computationally and neuroanatomically explicit account of the network of brain regions involved in speech acquisition and production. An overview of the model is provided along with descriptions of the computations performed in the different brain regions represented in the model. The latest version of the model, which contains a new right-lateralised feedback control map in ventral premotor cortex, will be described, and experimental results that motivated this new model component will be discussed. Application of the model to the study and treatment of communication disorders will also be briefly described.

Keywords: Speech production; DIVA model; Functional magnetic resonance imaging; Motor learning; Computational modeling.

INTRODUCTION

With the proliferation of functional brain imaging studies, a consensus regarding the brain areas underlying speech motor control is building (e.g.,
Indefrey & Levelt, 2004; Turkeltaub, Eden, Jones, & Zeffiro, 2002). Bohland and Guenther (2006) have described a ‘minimal network’ of brain regions involved in speech production that includes bilateral medial and lateral frontal cortex, parietal cortex, superior temporal cortex, the thalamus, basal ganglia, and the cerebellum. It is little surprise that these regions are commonly associated with the planning and execution of movements (primary sensorimotor and premotor cortex, the supplementary motor area, the cerebellum, thalamus, and basal ganglia) and those associated with acoustic and phonological processing of speech sounds (the superior temporal gyrus). A complete, mechanistic account of the role played by each region during speech production and how they interact to produce fluent speech is still lacking.

The goal of our research programme over the past 16 years has been to improve our understanding of the neural mechanisms that underlie speech motor control. Over that time we have developed a computational model of speech acquisition and production called the DIVA model (Guenther, 1994, 1995; Guenther, Ghosh, & Tourville, 2006; Guenther, Hampson, & Johnson, 1998). DIVA is an adaptive neural network that describes the sensorimotor interactions involved in articulator control during speech production. The model has been used to guide a number of behavioural and functional imaging studies of speech processing (e.g., Bohland & Guenther, 2006; Ghosh, Tourville & Guenther, 2008; Guenther et al., 1999; Lane et al., 2005, 2007a, 2007b; Nieto-Castanon, Guenther, Perkell, & Curtin, 2005; Perkell et al., 2004a, 2004b; Tourville, Reilly, & Guenther, 2008). The mathematically explicit nature of the model allows for straightforward comparisons of hypotheses generated from simulations of experimental conditions to empirical data. Simulations of the model generate predictions regarding the expected acoustic (e.g., formant frequencies), somatosensory (e.g., articulator positions), learning rates, and activity levels within specific model components. Experiments are designed to test these predictions, and the empirical findings are, in turn, used to further refine the model.

In its current form, the DIVA model provides a unified explanation of a number of speech production phenomena including motor equivalence (variable articulator configurations that produce the same acoustic output), contextual variability, anticipatory and carryover coarticulation, velocity/distance relationships, speaking rate effects, and speaking skill acquisition and retention throughout development (e.g., Callan, Kent, Guenther, & Vorperian, 2000; Guenther, 1994, 1995; Guenther et al., 2006; Guenther et al., 1998; Nieto-Castanon et al., 2005). Because it can account for such a wide array of data, the DIVA model has provided the theoretical framework for a number of investigations of normal and disordered speech production. Predictions from the model have guided studies of the role of auditory feedback in normally hearing persons, deaf persons, and persons who have recently regained some hearing through the use of cochlear implants (Lane et al., 2007a; Perkell et al.,
2000, 2004a, 2004b, 2007). The model has also been employed in investigations of the etiology of stuttering (Max, Guenther, Gracco, Ghosh, & Wallace, 2004), and acquired apraxia of speech (Robin et al., 2008; Terband, Maassen, Brumberg, & Guenther, 2008).

In this review, the key concepts of the DIVA model are described with a focus on recent modifications to the model. Our investigations of the brain regions involved in feedback-based articulator control have motivated the addition of a lateralised feedback control map in ventral premotor cortex of the right hemisphere. Additional brain regions known to contribute to speech motor control have also been incorporated in the model. Projections originating in the supplementary motor area and passing through the basal ganglia and thalamus are hypothesised to serve as gates on the outflow of motor commands. Support for these modifications and the impact they have on the model are discussed below.

DIVA MODEL OVERVIEW

The DIVA model, schematised in Figure 1, consists of integrated feedforward and feedback control subsystems. Together, they learn to control a simulated vocal tract, a modified version of the synthesiser described by Maeda (1990). Once trained, the model takes a speech sound as input, and generates a time varying sequence of articulator positions that command movements of the simulated vocal tract that produce the desired sound. Each block in Figure 1 corresponds to a set of neurons that constitute a neural representation. When describing the model, the term map is used to refer to such a set of cells, represented by boxes in Figure 1. The term mapping is used to refer to a transformation from one neural representation to another. These transformations are represented by arrows in Figure 1 and are assumed to be carried out by filtering cell activations in one map through synapses projecting to another map. The synaptic weights are learned during a babbling phase meant to coarsely represent that typically experienced by a normally developing infant (e.g., Oller & Eilers, 1988). Random movements of the speech articulators provide tactile, proprioceptive, and auditory feedback signals that are used to learn the mappings between the different neural representations. After babbling, the model can quickly learn to produce new sounds from audio samples provided to it.

An important part of the model’s development has been assigning components of the model to corresponding regions of the brain (see anatomical labels in boxes in Figure 1). Model components were mapped to locations in the Montreal Neurological Institute (MNI; Mazziotta et al., 2001) standard reference frame based on relevant neuroanatomical and neurophysiological studies (Guenther et al., 2006). The assignment of component
locations is based on the synthesis of a large body of behavioural, neurophysiological, lesion, and neuroanatomical data. A majority of these data were derived from studies focused specifically on speech processes; however, studies of other modalities (e.g., non-orofacial motor control) also contributed. Associating model components to brain regions allows (i) the generation of neuroanatomically specified hypotheses regarding the neural processes underlying speech motor control from a unified theoretical framework, and (ii) a comparison of model dynamics with past, present, and future clinical and physiological findings regarding the functional neuroanatomy of speech processes.

Blood oxygenation level-dependent (BOLD) functional magnetic resonance imaging (fMRI; Belliveau et al., 1992; Kwong et al., 1992; Ogawa et al., 1993) has provided a powerful tool for the non-invasive study of human brain function. BOLD signal provides an indirect measure of neural activity. The coupling between neural activity and changes in local blood oxygen levels

Figure 1. The DIVA model of speech acquisition and production. Recently added modules and connections are highlighted by black outlines. Model components are associated with hypothesized neuroanatomical substrates. Abbreviations: GP = globus pallidus; HG = Heschl’s gyrus; pIFg = posterior inferior frontal gyrus; pSTg = posterior superior temporal gyrus; Put = putamen; slCB = superior lateral cerebellum; smCB = superior medial thalamus; SMA = supplementary motor area; SMG = supramarginal gyrus; VA = ventral anterior nucleus of the thalamus; VL = ventral lateral nucleus of the thalamus; vMC = ventral motor cortex; vPMC = ventral premotor cortex; vSC = ventral somatosensory cortex. [To view this figure in colour, please visit the online version of this Journal.]
remains a matter of debate. While counter arguments have been made (Mukamel, Gelbard, Arieli, Hasson, Fried, & Malach, 2005), consensus is building around the hypotheses that BOLD signal is correlated with local field potentials (Goense & Logothetis, 2008; Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001; Mathiesen, Caesar, Akgoren, & Lauritzen, 1998; Viswanathan & Freeman, 2007). Local field potentials are thought to reflect local synaptic activity, i.e., a weighted sum of the inputs to a given region (Raichle & Mintun, 2006).

The development of BOLD fMRI has proven particularly beneficial to the study of speech given its uniquely human nature. The past decade and a half of imaging research has provided a tremendous amount of functional data regarding the brain regions involved in speech (e.g., Fiez & Petersen, 1998; Ghosh, Bohland, & Guenther, 2003; Indefrey & Levelt, 2004; Riecker et al., 2000b; Soros, Sokoloff, Bose, McIntosh, Graham, & Stuss, 2006; Turkeltaub et al., 2002), including the identification of a ‘minimal network’ of the brain regions involved in speech production (Bohland & Guenther, 2006). A clear picture of the contributions made by each region, and how these regions interact during speech production remains elusive, however.

We believe that the continued study of the neural mechanisms of speech will benefit from the combined use of functional neuroimaging and computational modelling. For this purpose, the proposed locations of model components have been mapped to anatomical landmarks of the canonical brain provided with the SPM image analysis software package (Friston et al., 1995; http://www.fil.ion.ucl.ac.uk/spm/). The SPM canonical brain is a popular substrate for presenting neuroimaging data. As such, it provides a familiar means of reference within MNI reference space. Mapping the model’s components onto this reference, then, provides a convenient means to compare the results of a large pool of neuroimaging experiments from a common theoretical framework, a framework that accounts for a wide range of data from diverse experimental modalities. In other words, it constrains the interpretation of fMRI results with classical lesion data, microstimulation findings, previous functional imaging work, etc. Mapping the model’s components to brain anatomy also permits the generation of anatomically explicit simulated haemodynamic responses based on the model’s cell activities. These predictions can then be used to constrain the design and interpretation of functional imaging studies (e.g., Ghosh et al., 2008; Tourville et al., 2008).

Guenther et al. (2006) detailed the neuroanatomical mapping of the model, including a discussion of evidence supporting the assignment of each location. Here we focus on additional assignments given recent modifications of the DIVA model. The MNI locations for recently added regions encompassed by the model are listed in Table 1. These sites are also plotted on a rendering of the SPM canonical brain surface in Figure 2.
TABLE 1
The location of new DIVA model components in MNI space

<table>
<thead>
<tr>
<th>Model components</th>
<th>Left hemisphere</th>
<th>Right hemisphere</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feedback Control Map</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right ventral premotor cortex</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Initiation Map</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMA</td>
<td>0 0 68</td>
<td>2 4 62</td>
</tr>
<tr>
<td>Putamen</td>
<td>-26 -2 4</td>
<td>30 -14 4</td>
</tr>
<tr>
<td>Globus Pallidus</td>
<td>-24 -2 -4</td>
<td>24 2 -2</td>
</tr>
<tr>
<td>Thalamus</td>
<td>-10 -14 8</td>
<td>10 -14 8</td>
</tr>
<tr>
<td><strong>Articulator Velocity and Position Maps</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larynx (Intrinsic)</td>
<td>-53 0 42</td>
<td>53 4 42</td>
</tr>
<tr>
<td>Larynx (Extrinsic)</td>
<td>-58.1 6.0 6.4</td>
<td>65.4 5.2 10.4</td>
</tr>
<tr>
<td><strong>Somatosensory State Map</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larynx (Intrinsic)</td>
<td>-53 -8 42</td>
<td>53 -14 38</td>
</tr>
<tr>
<td>Larynx (Extrinsic)</td>
<td>-61.8 1 7.5</td>
<td>65.4 1.2 12</td>
</tr>
</tbody>
</table>

Figure 2. Neuroanatomical mapping of the DIVA model. The location of DIVA model component sites (light grey dots) are plotted on renderings of the left and right lateral surfaces of the SPM2 canonical brain. Sites immediately anterior to the central sulcus (dotted line) represent cells of the model’s articulator velocity (M) and position (M) maps. Sites located immediately posterior to the central sulcus represent cells of the somatosensory state map (S). Subcortical sites (basal ganglia, thalamus, paravermal cerebellum, deep cerebellar nuclei), are not shown. Additional abbreviations: Au = auditory state map; ΔAu = auditory error map; FBM = feedback control map; IM = initiation map; Lat Cbm = lateral cerebellum; Resp = respiratory motor cells; ΔS = somatosensory error map; SSM = speech sound map; T\(_{Au}\) = auditory target map; T\(_{S}\) = somatosensory target map. [To view this figure in colour (where the component sites are shown as red dots), please visit the online version of this Journal.]
Feedforward control

Speech production begins in the model with the activation of a speech sound map cell in left premotor and adjacent inferior frontal cortex. According to the model, each frequently encountered speech sound in a speaker’s environment is represented by a unique cell in the speech sound map. Cells in the speech sound map project to cells in feedforward articulator velocity maps (labelled $\dot{M}$ in Figure 2) in bilateral ventral motor cortex. These projections represent the set of feedforward motor commands or articulatory gestures (cf. Browman & Goldstein, 1989) for that speech sound. The feedforward articulator velocity map in each hemisphere consists of eight antagonistic pairs of cells that encode movement velocities for the upper and lower lips, the jaw, the tongue, and the larynx. These velocities ultimately determine the positions of the eight articulators of the Maeda (1990) synthesiser (see section on articulator movements below for a description of this process). An active speech sound map cell sends a time-varying 16-dimensional input to the feedforward articulator velocity map that encodes the articulator velocities for production of a learned speech sound. The weights are learned during an imitation phase (see below). Feedforward articulator velocity maps are hypothesised to be distributed along the caudal portion of ventrolateral precentral gyrus, the region of primary motor cortex that controls movements of the speech articulators. These cells are hypothesised to correspond to ‘phasic’ cells that have been identified in recordings from motor cortex cells in monkeys (e.g., Kalaska, Cohen, Hyde, & Prud’homme, 1989). Ipsilateral and contralateral premotor-to-motor projections that would underlie this hypothesised connectivity have been demonstrated in monkeys (e.g., Dancause et al., 2006, 2007; Fang, Stepniewska, & Kaas, 2005; Stepniewska, Preuss, & Kaas, 2006). Modulation of primary motor cortex by ventral premotor cortex before and during movements has been shown in a number of studies (e.g., Cattaneo et al., 2005; Davare, Lemon, & Olivier, 2008). We expect that, in addition to direct premotor to primary motor cortex projections, additional projections via the basal ganglia and/or cerebellum (projecting back to cortex via the thalamus) are also involved in representing the feedforward motor programs.

The mapping from a cell in the speech sound map to the articulator velocity cells is analogous to the process of ‘phonetic encoding’ as conceptualised by Levelt et al. and colleagues (e.g., Levelt, Roelofs, & Meyer, 1999; Levelt & Wheeldon, 1994); i.e., it transforms a phonological input from adjacent inferior frontal cortex into the set of feedforward motor commands that produce that sound. The speech sound map, then, can be likened to Levelt et al.’s ‘mental syllabary’, a repository of learned speech motor programs. However, rather than being limited to a repository of motor programs for frequently produced syllables, as Levelt proposed, the speech
sound map also represents common syllabic, sub-syllabic (phonemes), and multi-syllabic speech sounds (e.g., words, phrases).

As mentioned above, we hypothesise that this repository of speech motor programs is located in the left hemisphere in right-handed speakers. A dominant role for the left hemisphere has been a hallmark of language processing models for several decades (Geschwind, 1970). Broca’s early findings linking left inferior frontal damage with speech production deficits have been corroborated by a large body of lesion studies (Dronkers, 1996; Duffy, 2005; Hillis et al., 2004; Kent & Tjaden, 1997). These clinical findings suggest that speech motor planning is predominantly reliant upon contribution from the posterior inferior frontal region of the left hemisphere. However, the specific level at which the production process becomes lateralised (e.g., semantic, syntactic, phonological, articulatory) has been a topic of debate. We recently showed that the production of single nonsense monosyllables, devoid of semantic or syntactic content, involves left-lateralised contributions from inferior frontal cortex including inferior frontal gyrus, pars opercularis (BA 44), ventral premotor, and ventral motor cortex (Ghosh et al., 2008). Activation in these areas is also left-lateralised during monosyllable word production (Tourville et al., 2008). These findings are consistent with the model’s assertion that feedforward articulator control originates from cells representing speech motor programs that lie in the left hemisphere.¹ This conclusion, based on neuroimaging data, is consistent with the classical left-hemisphere-dominance view that arose from lesion data. It implies that damage to the left inferior frontal cortex is more commonly associated with speech disruptions than damage to the same region in the right hemisphere because feedforward motor programs are disrupted.

This interpretation is relevant for the study and treatment of speech disorders such as acquired apraxia of speech (AOS). Lesions associated with AOS are predominantly located in the left hemisphere (Duffy, 2005), and particularly affect ventral BA 6 and 44 (ventral precentral gyrus, posterior inferior frontal gyrus, frontal operculum) and the underlying white matter. Our findings corroborate characterisations of AOS as a disruption of the use and development of speech motor programs (Ballard, Granier, & Robin, 2000; McNeil, Robin, & Schmidt, 2007) and suggest that rehabilitative treatments focused on restoring motor programs, e.g., sound production treatment (Wambaugh, Duffy, McNeil, Robin, & Rogers, 2006), and/or improving feedback-based performance should be emphasised.

¹ Performance of these motor programs relies on contributions from bilateral premotor, motor, and subcortical regions (see below), areas that are active bilaterally during normal speech production.
Feedback control

As indicated in Figure 1, the speech sound map is hypothesised to contribute to both feedforward and feedback control processes. In addition to its projections to the feedforward control map, the speech sound map also projects to auditory and somatosensory target maps. These projections encode the time-varying sensory expectations, or targets, associated with the active speech sound map cell. Auditory targets are given by three pairs of inputs to the auditory target map that describe the upper and lower bounds for the 1st, 2nd, and 3rd formant frequencies of the speech sound being produced. Somatosensory targets consist of a 22-dimensional vector that describes the expected proprioceptive and tactile feedback for the sound being produced. Projections such as these, which predict the sensorimotor state resulting from a movement, are typically described as representing a forward model of the movement (e.g., Davidson & Wolpert, 2005; Desmurget & Grafton, 2000; Kawato, 1999; Miall & Wolpert, 1996). According to the model, the auditory and somatosensory target maps send inhibitory inputs to auditory and somatosensory error maps, respectively. The error maps are effectively the inverse of the target maps: input to the target maps results in inhibition of the region of the error map that represents the expected sensory feedback for the sound being produced. Auditory target and error maps are currently hypothesised to lie in two locations along the posterior superior temporal gyrus. These sites, a lateral one near the superior temporal sulcus, and a medial one at the junction of the temporal and parietal lobes deep in the Sylvian fissure, respond both during speech perception and speech production (Buchsbaum, Hickok, & Humphries, 2001; Hickok & Poeppel, 2004). In the model, both sites are bilateral. The somatosensory target and state maps lie in ventral supramarginal gyrus, a region Hickok and colleagues (e.g., Hickok & Poeppel, 2004) have argued supports the integration of speech motor commands and sensory feedback. This hypothesised role for ventral parietal cortex during speech production is analogous to the visual-motor integration role associated with more dorsal parietal regions during limb movements (Andersen, 1997; Rizzolatti, Fogassi, & Gallese, 1997).

The sensory error maps also receive excitatory inputs from sensory state maps in auditory and somatosensory cortex. The auditory state map is hypothesised to lie along Heschl’s gyrus and adjacent anterior planum temporale, a region associated with primary and secondary auditory cortex. Cells in the somatosensory state map are distributed along the ventral precentral gyrus, roughly mirroring the motor representations on the opposite bank of the central sulcus (see Figure 2). Projections from the auditory and somatosensory state maps relay an estimate of the current sensory state. Activity in the error maps, then, represents the difference between the expected
and actual sensory states associated with the production of the current speech sound production.

By effectively cancelling the self-produced portion of the sensory feedback response, the speech sound map inputs to the sensory target maps function similarly to projections originally described by von Holst and Mittelstaedt (1950) and Sperry (1950). Von Holst and Mittelstaedt (1950) proposed the ‘principle of reafference’ in which a copy of the expected sensory consequences of a motor command, termed an efference copy, was subtracted from the realised sensory consequences. A wide body of evidence suggests such a mechanism plays an important role in the motor control of eye and hand movements, as well as speech (e.g., Bays, Flanagan, & Wolpert, 2006; Cullen, 2004; Heinks-Maldonado & Houde, 2005; Reppas, Usrey, & Reid, 2002; Roy & Cullen, 2004; Voss, Ingram, Haggard, & Wolpert, 2006). The hypothesised inhibition of higher order auditory cortex during speech production is supported by several recent studies. Wise and colleagues, using positron emission tomography (PET) to indirectly assess neural activity, noted reduced superior temporal gyrus activation during speech production compared with a listening task (Wise, Greene, Buchel, & Scott, 1999). Similarly, comparisons of auditory responses during self-produced speech and while listening to recordings of one’s own speech indicate attenuation of auditory cortex responses during speech production (Curio, Neuho, Numminen, Jousmaki, & Hari, 2000; Heinks-Maldonado, Mathalon, Gray, & Ford, 2005; Heinks-Maldonado, Nagarajan, & Houde, 2006; Numminen, Salmelin, & Hari, 1999). Further evidence of auditory response suppression during self-initiated vocalisations is provided by single unit recordings from non-human primates; for example, attenuation of auditory cortical responses prior to self-initiated vocalisations has been demonstrated in the marmoset (Eliades & Wang, 2003, 2005).

If incoming sensory feedback does not fall within the expected target region, an error signal is sent to the feedback control map in right frontal/ventral premotor cortex. The feedback control map transforms the auditory and somatosensory error signals into corrective motor velocity commands via projections to the articulator velocity map in bilateral motor cortex. The model’s name, DIVA, is an acronym for this mapping from sensory directions into velocities of articulators. Feedback-based articulator velocity commands are integrated and combined with the feedforward velocity commands in the articulator position map (see section on articulator movements below).

The right-lateralised feedback control map was added to the model based on recent results from neuroimaging investigations designed to reveal the neural substrates underlying feedback control of speech production. These studies used fMRI to compare brain activity during speech production under normal and perturbed auditory (Tourville et al., 2008) and somatosensory (Gollinopoulos, Tourville, Bohland, Ghosh, & Guenther, 2010) feedback
conditions. Left-lateralised activity in the posterior inferior frontal gyrus *par opercularis*, ventral premotor, and ventral primary motor cortex was noted during the normal feedback condition in both studies. When auditory feedback was perturbed, activity increased bilaterally in the posterior superior temporal cortex, the hypothesised location of the auditory error map, and speakers produced compensatory movements (evident by changes in the acoustic signals produced by the subjects in the perturbed condition compared with the unperturbed condition). The compensatory movements were associated with a right lateralised increase in ventral premotor activity. Structural equation modelling (see Tourville et al., 2008 for details) was used to investigate effective connectivity within the network of regions that contributed to feedback-based auditory control. This analysis revealed increased effective connectivity from left posterior temporal cortex to right posterior temporal and ventral premotor cortex (Figure 3). Evidence that right posterior temporal cortex exerts additional influence over motor output via a connection through right inferior frontal gyrus, *pars triangularis* (BA 45) during feedback control was also found.

Figure 3. Effective connectivity within the auditory feedback control network. Structural equation modeling demonstrated significant modulation of interregional interactions within the schematized network when auditory feedback was perturbed during speech production. Pairwise comparisons of path coefficients in the normal and perturbed feedback conditions revealed significant increases in the positive weights from left posterior superior temporal gyrus (pSTg) to right pSTg (the path labeled a in the diagram above), from left pSTg to right ventral premotor cortex (PMC; path b), and from right pSTg to right inferior frontal gyrus, *pars triangularis* (path c) when auditory feedback was perturbed during speech production. Additional abbreviation: MC = motor cortex.
Other imaging studies of speech production that have included a perturbed auditory feedback condition have demonstrated greater right hemisphere involvement in auditory feedback-based control of speech (e.g., Fu et al., 2006; Toyomura et al., 2007). We also noted the same right-lateralised increase of ventral premotor activity associated with auditory feedback perturbation when somatosensory feedback was perturbed (Golfinopoulos et al., 2010). A recent study of visuo-motor control reached similar conclusions regarding the relative contributions of the two hemispheres during feedforward- and feedback-based motor control (Grafton, Schmitt, Van Horn, & Diedrichsen, 2008). Thus, there is mounting evidence that these hemispheric differences may be a general property of the motor control system.

The implications of lateralised feedforward and feedback motor control of speech may be relevant to the study and treatment of stuttering. Neuroimaging studies of speech production in persons who stutter consistently demonstrate increased right hemisphere activation relative to normal speakers in the precentral and inferior frontal gyrus regions (see Brown, Ingham, Ingham, Laird, & Fox, 2005 for review), the same frontal regions identified as part of the feedback control network by Tourville et al. (2008). It has been hypothesised that stuttering involves excessive reliance upon auditory feedback control due to poor feedforward commands (Max et al., 2004). The findings of Tourville et al. provide support for this hypothesis: auditory feedback control during the perturbed feedback condition, clearly demonstrated by the behavioural results, was associated with increased activation of right precentral and inferior frontal cortex. According to this view, the right hemisphere inferior frontal activation is a secondary consequence of the root problem, which is aberrant performance in the feedforward system. Poor feedforward performance leads to auditory errors that in turn activate the right-lateralised auditory feedback control system in an attempt to correct for the errors. This hypothesis is consistent with the effects of fluency-inducing therapy on BOLD responses; successful treatment has been associated with a shift toward more normal, left-lateralised frontal activation (De Nil, Kroll, Lafaille, & Houle, 2003; Neumann et al., 2005).

Articulator movement

The feedforward velocity commands and feedback-based error corrective commands are integrated in the articulator position maps (labelled M in Figure 2) that lie along caudoventral precentral gyrus, adjacent to the feedforward articulator velocity maps. This area is the primary motor representation for muscles of the face and vocal tract. Cells in the articulator position maps are hypothesised to correspond to ‘tonic’ neurons that have been identified in monkey primary motor cortex (e.g., Kalaska et al., 1989).
The map consists of 10 pairs of antagonistic cells that correspond to parameters of the Maeda vocal tract that determine lip protrusion, upper and lower lip height, jaw height, tongue height, tongue shape, tongue body position, tongue tip location, larynx height, and glottal opening and pressure. Activity in the articulator position maps is a weighted sum of the inputs from the feedforward and feedback-based velocity commands. The relative weight of feedforward and feedback commands in the overall motor command is dependent upon the size of the error signal since this determines the size of the feedback control contribution. The resulting articulator position command drives the simulated vocal tract to produce the desired speech sound.

Based on recent imaging work (Brown, Ngan, & Liotti, 2008; Olthoff, Baudewig, Kruse, & Dechent, 2008), additional motor cortex cells have been added to the model that represent the intrinsic laryngeal muscles. The locus of a laryngeal representation in motor cortex has typically been associated with the ventrolateral extreme of the precentral gyrus (e.g., Duffy, 2005; Ludlow, 2005), an assumption based heavily upon findings from non-human primates (e.g. Simonyan & Jurgens, 2003) and supported by the intracortical mapping studies performed by Penfield and colleagues in humans prior to epilepsy surgery (Penfield & Rasmussen, 1950; Penfield & Roberts, 1959). Accordingly, a motor larynx representation in this location was included in our initial anatomical mapping of the model (Guenther et al., 2006). Brown et al. (2008) and Olthoff et al. (2008) have since demonstrated a bilateral representation in a more dorsal region of motor cortex adjacent to the lip area and near a second ‘vocalisation’ region identified by Penfield and Roberts (1959, p. 200). The authors also noted a ventral representation near/within bilateral Rolandic operculum that is consistent with the non-human primate literature. Brown and colleagues (2008) concluded that the dorsal region likely represents the intrinsic laryngeal muscles that control the size of the glottal opening. The opercular representation, it was speculated, likely represents the extrinsic laryngeal muscles that affect vocal tract resonances by controlling larynx height. Based on these findings, two sets of cells representing laryngeal parameters of the Maeda articulator model (Maeda, 1990) associated with laryngeal functions have been assigned to MNI locations: cells in ventrolateral precentral gyrus (labelled Larynx, Extrinsic in Table 1) represent larynx height, whereas cells in the dorso medial orofacial region of precentral gyrus (labelled Larynx, Intrinsic in Table 1) represent a weighted sum of parameters representing glottal opening and glottal pressure.

Speech production is consistently associated with bilateral activity in the medial prefrontal cortex, including the supplementary motor area (SMA), in 2

---

2 The activity of cells representing glottal opening and pressure are not determined dynamically via input from the feedforward and feedback velocity commands. Their activity is given by a weighted sum of Maeda vocal tract parameters that are set manually.
the basal ganglia, and in the thalamus (e.g., Bohland & Guenther, 2006; Ghosh et al., 2008; Tourville et al., 2008). Previous versions of the DIVA model have offered no account of this activity. The SMA is strongly interconnected with lateral motor and premotor cortex and the basal ganglia (Jurgens, 1984; Lehericy et al., 2004; Luppino, Matelli, Camarda, & Rizzolatti, 1993; Matsumoto et al., 2004, 2007). Recordings from primates have revealed cells in the SMA that encode movement dynamics (Padoa-Schioppa, Li, & Bizzi, 2004). Cells representing higher-order information regarding the planning/performance of sequences of movements have also been identified in the SMA. Activity representing particular sequences of movements to be performed (Shima & Tanji, 2000), intervals between specific movements within a sequence (Shima & Tanji, 2000), the ordinal position of movements within a sequence (Clower & Alexander, 1998), and the number of items remaining in a sequence (Sohn & Lee, 2007) have been noted in recordings of neurons in the SMA. Microstimulation of the SMA in humans yields vocalisation, word or syllable repetitions, and/or speech arrest (Penfield & Welch, 1951). Bilateral damage to these areas results in speech production deficits including transcortical motor aphasia (Jonas, 1981; Ziegler, Kilian, & Deger, 1997) and akinetic mutism (Adams, 1989, 1990; Nemeth, Hegedus, & Molnar, 1988). These data and recent imaging findings have led a number of investigators to conclude that the SMA plays a critical role in controlling the initiation of speech motor commands (e.g., Alario, Chainay, Lehericy, & Cohen, 2006; Bohland & Guenther, 2006; Jonas, 1987; Ziegler et al., 1997).

The SMA is reciprocally connected with the basal ganglia, another region widely believed to contribute to gating motor commands (e.g., Albin, Young, & Penney, 1995; Pickett, Kuniholm, Protopapas, Friedman, & Lieberman, 1998; Van Buren, 1963). The basal ganglia receive afferents from most areas of the cerebral cortex, including motor and prefrontal regions and, notably, associative and limbic cortices. Thus, the basal ganglia are well-suited for integrating contextual cues for the purpose of gating motor commands.

Based on these findings, an *initiation map*, hypothesised to lie in the SMA, has been added to the DIVA model. The initiation map gates the release articulator position commands to the periphery. According to the model, each speech motor program in the speech sound map is associated with a cell in the initiation map. Motor commands associated with that program are released when the corresponding initiation map cell becomes active. Activity in the initiation map ($I$) is given by:

$$I_i(t) = \begin{cases} 1 & \text{if the } i^{th} \text{ sound is being produced or perceived} \\ 0 & \text{otherwise} \end{cases}$$
The timing of initiation cell activity is governed by contextual inputs from the basal ganglia via the thalamus. Presently, this timing is simply based on a delayed input from the speech sound map. Cells representing the initiation map have been placed bilaterally in the SMA, caudate, putamen, globus pallidus, and thalamus, all of which demonstrate activity during simple speech production tasks (e.g., Ghosh et al., 2008; Tourville et al., 2008). A model of speech motor sequence planning and execution, the GODIVA model, has been developed (Bohland, Bullock, & Guenther, in press; Bohland & Guenther, 2006). The GODIVA model provides a comprehensive account of the interactions between the SMA, basal ganglia, motor, and premotor cortex that result in the gating of speech motor commands; due to space limitations we refer the interested reader to that publication for details.

LEARNING IN THE DIVA MODEL

Early babbling phase

Before DIVA is able to produce speech sounds, the mappings between the various components of the model must be learned. The model first learns the relationship between motor commands and their sensory consequences during a process analogous to infant babbling. The mappings that are tuned in this process are highlighted in the simplified DIVA block diagram shown in the left panel of Figure 4. For clarity, anatomical labels have been removed and the names of the model’s components have been shortened. During the babbling phase, pseudo-random articulator movements provide auditory and somatosensory feedback that is associated with the causative motor commands. The paired motor and sensory information is used to tune the synaptic projections from temporal and parietal sensory error maps to the feedforward control map. Once tuned, these projections transform sensory error signals into corrective motor velocity commands. This mapping from desired sensory outcome to the appropriate motor action is an inverse kinematic transformation and is often referred to as an inverse model (e.g., Kawato, 1999; Wolpert & Kawato, 1998).

The cerebellum is a likely contributor to the feedback motor command. Neuroimaging studies of motor learning have noted cerebellar activity that is associated with the size or frequency of sensory error (e.g., Blakemore, Frith, & Wolpert, 2001; Blakemore, Wolpert, & Frith, 1999; Diedrichsen, Hashambhoy, Rane, & Shadmehr, 2005; Flament, Ellermann, Kim, Ugurbil, & Ebner, 1996; Grafton et al., 2008; Imamizu et al., 2000; Imamizu, Higuchi, Toda, & Kawato, 2007; Miall & Jenkinson, 2005; Schreurs et al., 1997; Tesche & Karhu, 2000). It has been speculated that a representation of sensory errors in the cerebellum drives corrective motor commands (Grafton et al., 2008; Penhune & Doyon, 2005) and contributes to feedback-based motor learning.
For instance, the cerebellum has been hypothesised to support the learning of inverse kinematics (e.g., Kawato, 1999; Wolpert & Kawato, 1998), a role for which it is anatomically well-suited; the cerebellum receives inputs from higher-order auditory and somatosensory areas (e.g., Schmahmann & Pandya, 1997), and projects heavily to the motor cortex (Middleton & Strick, 1997). Based on the cerebellum’s putative role in feedback-based motor learning, it is hypothesised to contribute to the mapping between sensory states and motor cortex, i.e., the projections that encode the feedback motor command.

**Imitation phase**

*Feedback control system.* Once the general sensory-to-motor mapping described above has been learned, the model undergoes a second learning phase that is specific to the production of speech sounds (Figure 4, right panel). This phase can be subdivided into two components. In the first component, weights from the speech sound map to the auditory target map are tuned. Analogous to the exposure an infant has to the sounds of his/her native language, the model is presented speech sound samples (e.g., phonemes, syllables, words). The speech samples take the form of time varying acoustic signals spoken by a human speaker. According to the model, when a new
speech sound is presented, it becomes associated with an unused cell in the inferior frontal speech sound map (via temporo-frontal projections not shown in Figure 2). With subsequent exposures to that speech sound, the model learns an auditory target for that sound in the form of a time-varying region that encodes the allowable variability in the acoustic signal (see Guenther, 1995, for a description of the learning laws that govern this process). During the second component of learning in the feedback control system, weights from the speech sound map to the somatosensory target map are tuned during correct self productions.

Reciprocal pathways between inferior frontal and auditory and somatosensory cortex have been demonstrated in humans (Makris et al., 2005; Matsumoto et al., 2004) and non-human primates (Morel & Kaas, 1992; Ojemann, 1991; Romanski et al., 1999; Schmahmann & Pandya, 2006); also see Duffau (2008) for a description of the putative fronto-parietal and fronto-temporal pathways involved in language processing. It has been argued by many that the cerebellum uses sensory error to build forward models that generate sensory predictions (Blakemore et al., 2001; Imamizu et al., 2000; Kawato et al., 2003; O’Reilly, Mesulam, & Nobre, 2008), the role of projections from the speech sound map to sensory target maps in DIVA model. It is likely, therefore, that the cerebellum contributes to the attenuation of sensory target representation in sensory cortex (cf. Blakemore et al., 2001). For this reason, cerebellar side loops are hypothesised in the projections from the speech sound map to the sensory target maps.

**Feedforward control system.** Feedforward commands are also learned during the imitation phase, once auditory targets have been learned. Initial attempts to produce the speech sound result in large sensory error signals due to poorly tuned projections from the speech sound map cells to the primary motor cortex articulatory velocity and position maps, and production relies heavily on the feedback control system. With each production, however, the feedback-based corrective motor command is added to the weights from the speech sound map to the feedforward articulator velocity cells, incrementally improving the accuracy of the feedforward motor command. With practice, the feedforward commands become capable of driving production of the speech sound with minimal sensory error and, therefore, little reliance on the feedback control system unless unexpected sensory feedback is encountered (e.g., due to changing vocal tract dynamics, a bite block, or artificial auditory feedback perturbation).

It is widely held that the cerebellum is involved with the learning and maintenance of feedforward motor commands (though see Grafton et al., 2008; Kawato, 1999; Ohyama, Nores, Murphy, & Mauk, 2003). The cerebellum receives input from premotor, auditory, and somatosensory cortical areas via the pontine nuclei and projects heavily to the motor cortex.
via the ventral thalamus (Middleton & Strick, 1997). This circuitry provides a substrate for the integration of sensory state information that may be important for choosing motor commands (e.g., Schmahmann & Pandya, 1997). In the DIVA model, the cerebellum is therefore included as a side loop in the projection from the speech sound map to the articulator velocity map.

Lesions to anterior vermal and paravermal cerebellum have been associated with disruptions of speech production (Ackermann, Vogel, Petersen, & Poremba, 1992; Urban et al., 2003), termed ataxic dysarthria, characterised by an impaired ability to produce speech with fluent timing and gestural coordination. This region is typically active bilaterally during overt speech production in neuroimaging experiments. Additional activity is typically found in adjacent lateral cortex bilaterally (Bohland & Guenther, 2006; Ghosh et al., 2008; Riecker, Ackermann, Wildgruber, Dogil, & Grodd, 2000; Riecker, Wildgruber, Dogil, Grodd, & Ackermann, 2002; Tourville et al., 2008; Wildgruber, Ackermann, & Grodd, 2001), an area less commonly associated with ataxic dysarthria. Model cells have therefore been placed bilaterally in two cerebellar cortical regions: anterior paravermal cortex (not visible in Figure 2) and superior lateral cortex (Lat. Cbm). The former are part of the feedforward control system, while the latter are hypothesised to contribute to the sensory predictions that form the auditory and somatosensory targets in the feedback control system.

The speech sound map and mirror neurons

The role played by the speech sound map in the DIVA model is similar to that attributed to ‘mirror neurons’ (Kohler et al., 2002; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996), so termed because they respond both while performing an action and perceiving an action. Mirror neurons in non-human primates have been shown to code for complex actions such as grasping rather than the individual movements that comprise an action (Rizzolatti et al., 1988). Neurons within the speech sound map are hypothesised to embody similar properties: activation during speech production drives complex articulator movement via projections to articulator velocity cells in motor cortex, and activation during speech perception tunes connections between the speech sound map and sensory target maps in auditory and somatosensory cortex. Evidence of mirror neurons in humans has implicated left precentral gyrus for grasping actions (Tai, Scherfler, Brooks, Sawamoto, & Castiello, 2004), and left opercular inferior frontal gyrus for finger movements (Iacoboni et al., 1999). Mirror neurons related to communicative mouth movements have been found in monkey area F5 (Ferrari, Gallese, Rizzolatti, & Fogassi, 2003) immediately lateral to their location for grasping movements (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). This area has been proposed to correspond to the caudal portion of ventral inferior frontal gyrus (Brodmann’s...
area 44) in the human (see Binkofski & Buccino, 2004; Rizzolatti & Arbib, 1998).

CURRENT PERSPECTIVES

The DIVA model provides a computationally explicit account of the interactions between the brain regions involved in speech acquisition and production. The model has proven to be a valuable tool for studying the mechanisms underlying normal (Callan et al., 2000; Lane et al., 2007b; Perkell et al., 2000, 2004a, 2004b, 2007; Villacorta, Perkell, & Guenther, 2007) and disordered speech (Max et al., 2004; Robin et al., 2008; Terband et al., 2008). Because the model is expressed as a neural network, it provides a convenient substrate for generating predictions that are well-suited for empirical testing. Importantly, the model’s development has been constrained to biologically plausible mechanisms. Thus, as DIVA has come to account for a wide range of speech production phenomena (e.g., Callan et al., 2000; Guenther, 1994, 1995; Guenther et al., 1998; Nieto-Castanon et al., 2005), it does so from a unified quantitative and neurobiologically grounded framework.

In this article, we have reviewed the key elements of the DIVA model, focusing on recent developments based on results from functional imaging experiments. Feedforward and feedback control maps are hypothesised to lie in left and right ventral premotor cortex, respectively. The lateralised motor control mechanisms embodied by the model may provide useful insight into the study and treatment of speech disorders. Questions remain, however, regarding the interaction between lateralised frontal and largely bilateral sensory processes. As discussed above, the data are consistent with DIVA-predicted projections from left-lateralised premotor cells to bilateral auditory cortex that encode sensory expectations. A further prediction of the model is that clear and stressed speech involve the use of smaller, more precise sensory targets compared with normal or fast speech (Guenther, 1995). If correct, increased activity should be seen in auditory and somatosensory cortical areas during clear and stressed speaking conditions, corresponding to increased error cell activity due to the more precise sensory targets. This prediction is currently being tested in an ongoing fMRI experiment.

Our experimental data also suggest that projections from bilateral auditory cells to right premotor cortex are involved in transforming auditory errors into corrective motor commands. The anatomical pathways that support these mechanisms are not fully understood. Further study of the information conveyed by those projections is also necessary. Studies have begun to explore putative sensory expectation projections from lateral frontal cortex to auditory cortex, establishing an inhibitory effect linked to ongoing articulator movements (e.g., Heinks-Maldonado et al., 2006). A clear understanding of
the units of this inhibitory input (e.g., does it have an acoustic, articulatory, or phonological organisation), as well as the error maps themselves, is yet to be fully established. Functional imaging experiments focused on these questions are currently underway.

The model has also expanded to include representations of the supplementary motor area and basal ganglia, which are hypothesised to provide a gating signal that initiates the release of motor commands to the speech articulators. In its current form, this initiation map is highly simplified. Mechanisms for learning the appropriate timing of motor command release have not yet been incorporated into the model. The brain regions associated with the model’s initiation map, along with the pre-supplementary motor area (e.g., Clower & Alexander, 1998; Shima & Tanji, 2000), have also been implicated in the selection and proper sequencing of individual motor programs for serial production. Bohland and Guenther (2006) investigated the brain regions that contribute to the assembly and performance of speech sound sequences. A neural network model of the mechanisms underlying this process, including interactions between the various cortical and subcortical brain regions involved, has been developed (Bohland et al., in press). Outputs from this speech sequencing model, termed GODIVA, serve as inputs to the DIVA model’s speech sound map. This work bridges the gap between computational models of the linguistic/phonological level of speech production (Dell, 1986; Hartley & Houghton, 1996; Levelt et al., 1999), and DIVA, which addresses production at the speech motor control level. Like the DIVA model, GODIVA is neurobiologically plausible and thus represents an expanded substrate for the study of normal and disorder speech processing.

An important aspect of speech production not addressed in the DIVA model is the control of prosody. Modulation of pitch, loudness, duration, and rhythm convey meaningful linguistic and affective cues (Bolinger, 1961, 1989; Lehiste, 1970, 1976; Netsell, 1973; Shriberg & Kent, 1982). The DIVA model has addressed speech motor control at the segmental level (phoneme or syllable units). With the development of the GODIVA model, we have started to account for speech production at a suprasegmental level. We have recently begun a similar expansion of the model to allow for control of prosodic cues, which often operates over multiple individual segments. Experiments currently under way are investigating whether the various prosodic features (volume, duration, and pitch) are controlled independently to achieve a desired stress level or if a combined ‘stress target’ is set that is reached by a dynamic combination of individual features. We are testing these alternative hypotheses measuring speaker compensations to perturbations of pitch and loudness. Adaptive responses limited to the perturbed modality supports the notion that prosodic features are independently controlled; adaptation across modalities is evidence for an integrated ‘stress’ controller.
As we have done in the past, we intend to couple our modelling efforts with investigations of the neural bases of prosodic control. There is agreement in the literature that no single brain region is responsible for prosodic control, but there is little consensus regarding which regions are involved and in what capacity (see Sidtis & Van Lancker Sidtis, 2003, for review). One of the more consistent findings in the literature concerns perception and production of affective prosody, which appear to rely more on the right cerebral hemisphere than the left hemisphere (Adolphs, Damasio, & Tranel, 2002; Buchanan et al., 2000; George et al., 1996; Ghacibeh & Heilman, 2003; Kotz et al., 2003; Mitchell, Elliott, Barry, Cruttenden, & Woodruff, 2003; Pihan, Altenmuller, & Ackermann, 1997; Ross & Mesulam, 1979; Williamson, Harrison, Shenal, Rhodes, & Demaree, 2003), though the view of affective prosody as a unitary, purely right hemisphere entity is oversimplified (Sidtis & Van Lancker Sidtis, 2003) and there is considerable debate over which and to what extent prosodic feature control is lateralised (e.g., Doherty, West, Dilley, Shattuck-Hufnagel, & Caplan, 2004; Emmorey, 1987; Gandour et al., 2003; Meyer, Alter, Friederici, Lohmann, & von Cramon, 2002; Stiller et al., 1997; Walker, Pelletier, & Reif, 2004). We hope to clarify this understanding by comparing the neural responses associated with prosodic control with those involved in formant control as indicated by our formant perturbation imaging study (Tourville et al., 2008). Our focus is on differences in the two control networks, particularly the laterality of sensory and motor cortical response. With this effort we hope to continue our progress toward building a comprehensive, unified account of the neural mechanisms underlying speech motor control.

REFERENCES


