Speech as a Problem of Motor Control in Robotics

Michael Connolly Brady (mcbrady@indiana.edu)
Indiana University, Cognitive Science Program
838 Eigenmann, 1910 E. 10th St., Bloomington, IN 47406, USA

Abstract

The conventional approach to speech production assumes that a linguistic control signal feeds down into an execution module where vocal articulators are coordinated. The linguistic signal takes the form of a stream of phonological units or discrete symbolic commands. This characterization reflects how a variety of control architectures in cognitive robotics are also based on symbolic commands. There are problems with symbolic motor control and in robotics there are alternatives to the assumption of symbols. This paper focuses on one such alternative. A minimal neural field model for speech motor planning and production is introduced. The model illustrates how some simple words may be represented for perception and production without coding the words in terms of phonological units. Concluding discussion considers how a scaled version of the model supports a construction grammar account of speech and language.

Keywords: speech perception and action; cognitive robotics; articulatory phonology; construction grammar.

Introduction

In analyzing tongue twisters and spoonerisms, we see that the apparent production components of an utterance may sometimes interfere with one other. To explain a spoonerism under the conventional view, we might attribute the error to problems with the motor plan. Production units were somehow sent in the incorrect order. Alternatively, we might attribute the error to a production mechanism that for some reason confused its instructions. By isolating the speech plan from its production, we are obliged to accept that a word-swap error is *either* an error in planning *or* an error in production. When it comes to implementation there can be no ambiguous middle ground.

What are the implications of a planning-production dichotomy? Here is the dilemma. In supposing that speech unit swap errors are not planned (why would we plan errors), we are left to believe that swap errors result from troubles in production. Yet if the production process is responsible for the serial arrangement of the apparent components of motor output, a production module would need to be provided with simultaneous access to the multiple components of the motor plan. It follows that if a production module somehow operates on multiple control instructions concurrently, it is problematic to conceptualize the control signal as being serially structured. The dilemma is resolved by rejecting the assumption that the control signal for speech is a stream of phonological commands. Motor planning and production are integrated. Control comes from an abstract and persistent signal rather than from a sequence of symbolic units to be executed one by one as they arrive.

Modular vs. Integrated Planning-Production

The relatively new field of cognitive robotics already offers a variety of models for conceptualizing how complex motor sequence production may be achieved. In analyzing these models, a classification scheme quickly becomes apparent. Models tend to be either modular or integrated. For the sake of definition, a modular approach conceptualizes the motor plan and its execution as separate processing tasks, to be handled by separate modules. In contrast, an integrated approach views motor control to come from a general and relatively persistent signal where the details of motor output are handled by planning-production dynamics. Sophisticated cognitive robotics architectures are usually not well characterized as being of solely one or the other approach, but a brief discussion with some explicit examples is in order to help better appreciate this modular versus integrated distinction.

The Theory of Articulatory Phonology (Browman and Goldstein, 1992, Goldstein et. al., 2006, Goldstein et. al., 2007, Saltzman & Kelso, 1987) is popularly interpreted to reflect the modular view. A stream of phonological units arrives from a singular source to be executed by a task dynamic model. The task dynamic model coordinates vocal motors in terms of articulatory gestures. An articulatory gesture is a related set of vocalic motor movements, also referred to as an 'action primitive' or 'action unit.' Action primitives combine into 'molecules' that correspond to phonological commands. The task dynamic model is responsible for streaming together articulatory gestures to produce fluid speech. Articulatory Phonology preserves the concept of the phonologically structured mental lexicon and related theories of generative syntax by explaining the lack of invariance in the speech stream as the result of motor production dynamics. Figure 1 illustrates the popular view as it depicts speech units in terms of "Hockett's" Easter eggs (Hockett, 1955). It should be noted that the bulk of Articulatory Phonology is not in conflict with an integrated perspective, a topic this paper will return to. The point being made here is that the widely held premise of a phonological control signal arriving from an executive source reflects the modular approach.

Beer et. al. (1992) provides an example of the integrated perspective using a six-legged robot that walks like an insect. A Continuous Time Recurrent Neural Network (CTRNN)-based nervous system coordinates the robot's legs for walking via the interactive dynamics of leg positions. The phase of each leg informs the phase of other legs so that walking behavior emerges without a control signal feeding down to tell each leg specifically what to do. A single parameter specifies how fast the robot should

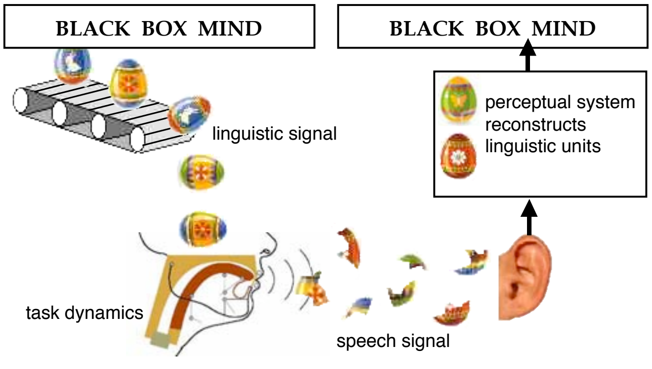

Figure 1: the classic view of speech

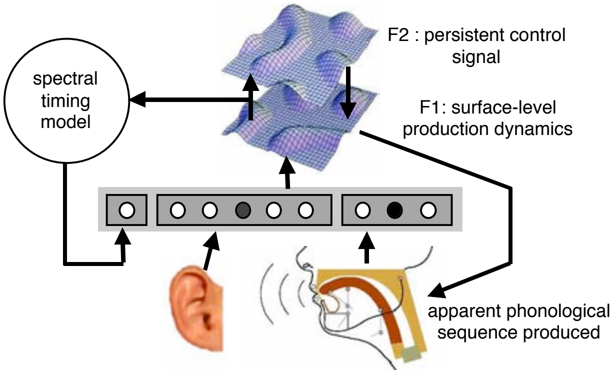
spectral timing model
F2 : persistent control signal
F1: surface-level production dynamics
apparent phonological sequence produced

Figure 2: this paper's alternative to the classic view

travel. By adjusting this speed parameter, analogous to turning the volume knob on a radio, the interactive leg dynamics of the robot shift to produce changes along a continuum in walking gait. Different gaits make the robot travel at different speeds. Here, the control signal or 'motor plan' is conceptualized simply as the speed control signal. The signal integrates with system dynamics to determine output motor behavior.

The task of programming a multi-joint robot arm to successfully reach for an object serves as a dual example for contrasting a modular with an integrated approach. With a modular approach, a vision system determines the position of an object in Cartesian space, calculates the set of joint angles that the robot arm would need to have in order to find the end of the arm at that position, and the arm is then instructed to move to attain those joint angle positions. Using this sense-plan-act method, an execution module interprets the command and smoothly moves the arm to the target. An integrated approach handles the task in a very different way. A mechanism is built where the arm moves to the target position using a feedback loop where the distance between the end of the arm and the target is systematically reduced to zero through time. Required joint angle positions of the arm are never explicitly calculated and the only control signal comes in the form of persistent parameters, such as the speed at which the arm should travel. E.g. see (Hersch & Billard, 2006). For further discussion on this general topic, the reader is directed to review the Equilibrium Point Hypothesis (EPH) in motor control.

The purpose of the model presented in the following pages is to illustrate how the integrated view may be implemented for speech. As with an integrated approach to robotic arm reaching and with Beer's insect crawler, explicit instructions from an executive controller are not found. Rather, volitional control comes from a persistent signal that influences production dynamics so that motor output behavior is realized through complex interactive processes.

The Model

Figure 2 introduces the model. It is constructed of dynamic fields that interact with each other through adaptive weights. Input is acoustic sound and motor feedback while output from the model controls an articulatory speech synthesizer.

A Basic Field

The model is founded on dynamic fields. A field is a two dimensional array of units where each unit is updated once per 5 millisecond time step with the equation:

$$\dot{u}_i = -u_i + S_i + h_f - \phi_i + n + \sum_j w_{ij} \cdot \sigma(u_j) \qquad (1)$$

The change in activation of a unit, *u*, is determined by the sum of influence to the unit minus its current activation. This influence comes from an outside signal, *S*, the field's slightly negative resting bias, *h*, a fatigue term for the unit, ϕ, a noise term, *n*, and from other units within the field. Fatigue for the unit increases as a function of time while the unit is active and decreases over time while the unit is inactive[1]. Influence from other units in the field is determined as the sum of the squashed[2] activations of neighboring units multiplied through corresponding within-field connection weights, *w*. These within field weights are specified by a Mexican hat function[3]. Input to the function is the Pythagorean distance between two units and output of the function specifies their connection weight.

A field is mathematically shaped like a torus so that all units have the same sized neighborhood of surrounding units. If given no outside input, *S,* and assuming a well-selected set of parameters[4], a randomly initialized field quickly approaches a non-zero, non-saturation equilibrium state. The contoured grid in Figure 3 illustrates a 30x30 field of units near such an equilibrium state after a number of iterations of Equation 1. Here we may assume that all units of the field were initialized with small random values and had received no input from outside the field. Due to the on-center off-surround nature of the Mexican hat, regions of the field that were initially slightly more active than other regions became very active to suppress regions that were initially only slightly less active.

A field is essentially a change detector and is 'tuned' to respond to specific patterns of change by adjusting the weights that carry input to the field. After training its weights, a field's equilibrium state deterministically reflects the change in input that has recently arrived to the field. To best introduce this, let us first walk through how input to a field is realized from the acoustic signal.

Acoustic Input

Raw sound is processed into frequency bands. The positive change in power in each band is input to the model via sensory nodes through time. Figure 3 depicts this as the vowel transition /i/→/a/ is processed. The spectrogram at the top of the figure shows the vowel transition and how it is split into five bands. The average power in each band through time is found. These band averages are illustrated in the diagram to the lower left with solid lines. Corresponding dashed lines in the diagram depict the positive change in the band-pass signals and this is what is provided as input to the network from the sensory nodes. At the snapshot in time depicted, there is virtually no change in the band-pass signals and thus the input nodes all have activations of zero. Now imagine as processing proceeds from left to right through the vowel transition. The input node corresponding to Band 3 experiences a momentary jump in activation (the power in other bands mostly does not change or changes negatively so that inputs for those bands remain at zero).

Node activations are passed to the field using a fully connected set of weights called an *adaptive filter*. During the transition between vowels of Figure 3, the equilibrium of the field is perturbed by the activation of Node 3. Depending on the values of the weights from Node 3 to the field, the field will be bumped out of its current equilibrium state towards a new equilibrium state.

Multiple Fields and Adaptive Filters

By adding a field to the system of Figure 3, we arrive at the minimal network depicted to the left in Figure 4. Fields are labeled F1 and F2 and adaptive filters are now depicted with singular arrows. The higher field (F2) responds to the lower field (F1) in the same way that F1 responds to sensory input. To provide this feedforward input to F2, F1 is sectioned into 25 zones. A zone’s value is found as a function of the sum of activation of a region of units, as depicted in the right diagram of Figure 4. The change in average activation of each zone through time provides feedforward input, analogous to a sensory node. Like with steady-state acoustic input, as long as there is not much change in a sending field's equilibrium state, there is no significant perturbation to the receiving field.

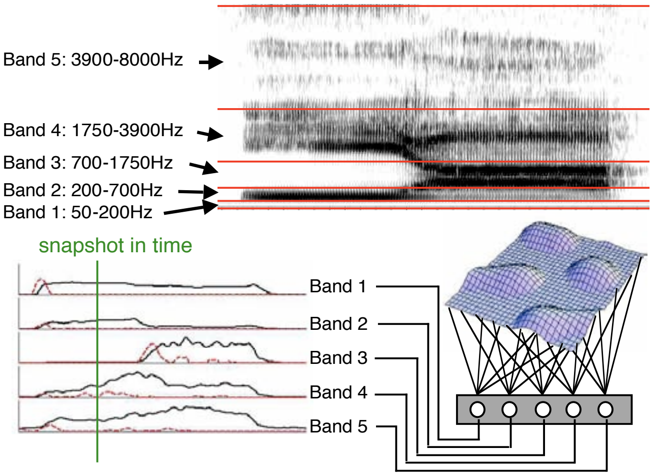

Figure 3: vowel transition /i/→/a/ as input to a field

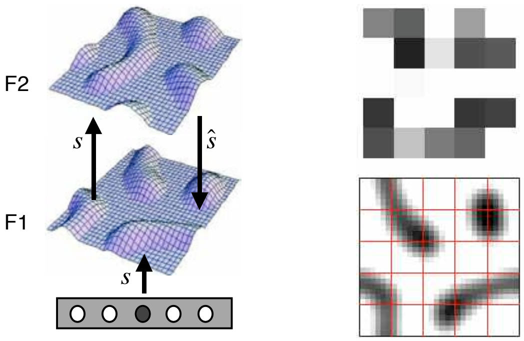

Figure 4: layering two fields into a minimal network (left); F1 is sectioned into zones for input to F2 (right)

Notice also in Figure 4 that there is an adaptive filter connecting F2 back to F1. This is a *feedback filter* and its purpose is to prime F1 for expected input. Like F1, F2 is divided into zones. These zones act as nodes to give feedback to F1 (and to send feedforward input to other fields). For feedback however, the zone or node value rather than the change in node value is what serves as input to the adaptive filter.

Equation 2 describes these between-field interactions. The input signal, S, to a unit in F1 is determined from the feedforward signal, s, to the unit and from the feedback signal, $\hat{s}$, to the unit. The feedforward signal is multiplied by a constant, k, to adjust the feedforward-feedback ratio. A gain term, g, is also provided. Feedback and feedforward signals are the squashed node values, o, or change in squashed node values, $\dot{o}$, passed through adaptive filters:

$$(2) \qquad S_i = g(ks_i + \hat{s}_i) \qquad s_i = \sum_j w_{ij}\dot{o}_j \qquad \hat{s}_i = \sum_j w_{ij}o_j$$

Let us step through a specific example of how processing works with the assumption that weights have already been trained. Consider that the equilibrium pattern of activation depicted on F1 in Figure 4 is the result of the /i/→/a/ vowel transition. Then consider another transition: /a/→/u/, yet to arrive. When this second transition arrives, F1 quickly moves toward a second equilibrium state corresponding to /au/. The transition between the two equilibrium patterns in F1, /ia/→/au/, is 'recognized' as F2 is perturbed toward a new equilibrium. This F2 equilibrium corresponds to the triphthong or word: /iau/, "~ yeow." Next assume that F2 was in its "yeow" equilibrium before input to the system began (the result of hypothetical feedback from other fields). Because F2 is already at the equilibrium it goes to, there is no change in F2's activation pattern. Thus, the persistant feedback signal from F2 'primes' F1 for both the /ia/ and /au/ transitions. In providing this top-down priming, the /au/ and /ia/ transitions are more readily detected by F1 as sensory input arrives. We will pursue this further in the next section, first it is important to introduce the remainder of the minimal model.

Motor Output and Proprioceptive Feedback

Changes in equilibrium state attractors of F1 correspond to coordinated motor movements in both perception and production. For the sake of simplicity, let us consider these motor movements in terms of category nodes with the understanding that activations of category nodes may be mapped to motor movements. Figure 5 depicts a set of three motor category nodes added to what we will now call the input gateway. Output from F1 to these motor nodes is like feedforward output from F1 to F2 where change in activation of a zone is the basis of the signal and category nodes are the simple sum of their input. When there is no change to the equilibrium state of F1, the motor nodes go to zero and there is no movement. The motor category nodes also provide proprioceptive input to F1 using feedback connections. The motor node array is thus analogous to F2 in its connectivity to F1. It is interesting to note however that connections from the motor category nodes to F1 can be considered as feedforward because the activations of the motor nodes correspond to change in motor parameters.

Timing Coordination

Timing is important in different ways for different languages. For example, in English the contrast between "fussy" and "fuzzy" relies on relative voice onset timing. In other languages, phonological distinctions may be based on segment durations. In Japanese one might accidentally call one's aunt (/obasan/) their grandmother (/obaasan/) simply by increasing the duration of the first /a/ vowel. A model for complex motor sequence planning and production requires a mechanism for encoding the timing relationships of the components in the sequence.

A spectral timing model is added to the minimal network of Figure 4 as depicted in Figure 2. It is called a spectral model because it is built of a bank of resonators, each tuned to resonate at a slightly different frequency. Resonators are mathematically described as pendulums (Brady, 2006), where an input value pushes the pendulum per time step. The same pushing is given to all resonators and a push is distributed across time. Think of short temporally patterned gusts of wind blowing the swings on a playground swing set where each swing has a different chain length. The input to a resonator at a given time step - analogous to a sample from a wind gust - is the sum of change in zone values of F1. Resonators with natural periods related to patterns of periodic activity in F1 achieve high amplitudes while resonators that do not relate to patterns of change in F1 do not achieve high amplitudes. As a resonator passes a 'firing phase,' (the middle of the pendulum's forward swing), it outputs a function of its amplitude to a summation node or timing 'expectancy' node. This node provides input through the input gateway back to F1 as depicted in Figure 5.

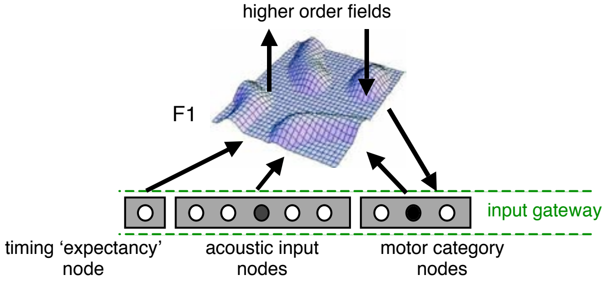

Figure 5: motor output and proprioceptive feedback is conceptualized in terms of motor category nodes

Figure 6 walks through the function of the spectral model as the network responds to the utterance: "got to be yeow." Track 1 of the figure is a notated spectrogram of this recorded utterance. Track 2 shows the activations of each of the five acoustic input nodes through time for the utterance. Track 3 depicts the sum of positive change of zones in F1 through time. This is the then the signal that is input to all resonators of the spectral model. Note that /bi/ perturbs F1 into an equilibrium state and F1 remains in that state through the /i/ vowel of the word "yeow." Track four presents the response of the spectral timing model (note: feedback from this node does not influence F1 in Figure 6). Lastly, Track 5 introduces a sinusoidal wave that depicts if the oscillator bank will emphasize or filter an input event based on the periodic structure of the pattern.

The role of the spectral model is to provide a reference signal for perception and motor coordination - effectively allowing the network to process events with respect to temporal structure. The response of the resonator bank to a given temporal pattern is deterministic regardless of whether the pattern exhibits periodic structure. Furthermore, a given temporal pattern generally elicits the same pattern of response from the resonator bank regardless of initial conditions. Lastly, the oscillator bank generalizes over rate. If an input pattern such as the one presented in Track 3 of Figure 6 were to unfold at a faster or slower rate, the oscillator bank would provide consistent values at the timing expectation node *relative* to events as they unfold. Imagine all five tracks of Figure 6 being horizontally compressed or stretched together.

As is noted in the next section, feedback from the spectral model is shown to influence segment duration, as with phonological distinctions in Japanese. For a discussion on timing analysis by oscillation see: (Brady & Port, 2007).

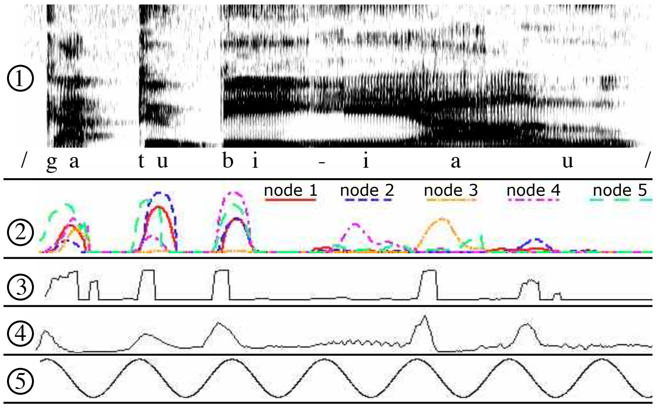

Figure 6: visualizing response of spectral timing model

Evaluation Scenario and Results

A simplistic evaluation scenario is pursued for the purpose of this paper. Three vowel transitions {/au/, /ia/, /ui/} were recorded from a male actor (this author). The frequency bands and change in power of the bands corresponding to sensory node input for /au/ and /ui/ are shown in the bottom of Figure 7 using the same frequency bands as depicted in Figure 3 for /ia/. Notice that for /au/, the activation of Auditory Node 2 is indicative of the transition. For /ia/, remember that Node 3 is the only node to become significantly active through the transition. And for /ui/, Auditory Node 4 is the node that cues the transition.

By allowing an isolated field to settle four times, each time from different random initial conditions (with fatigue and noise turned off), four static equilibrium targets were generated. Using the delta rule[5], each of these four targets was associated with a unique static input vector as depicted in the tops of Figure 7. Notice that the timing expectation node is on for the first three vectors and that the fourth vector (/aau/) is a version of the third (/au/) except with the timing expectation node turned off. After training, when one of the vowel transitions is input (with a corresponding timing node on/off value and with or without motor feedback), F1 is perturbed to quickly shift towards the transition's trained equilibrium target. Connections from F1 to the motor nodes were also trained in this manner.

Four more equilibrium state targets from random initial conditions were generated. These targets correspond to the triphthongs or words {/iau/, /uia/, /aui/ and /iaau/} and are depicted as the four static F2 patterns in Figure 8. Weights can be trained from F1 to F2, but that does not concern us now. In this analysis we need only to consider feedback from F2 to F1. Pairs of F1 activation patterns from Figure 7 were summed to provide targets for training the F2 feedback filter. For example, the equilibrium of F2 corresponding to the word /iau/ in Figure 8 is associated (again using the delta rule) with *both* the /ia/ and /au/ transitions because its static F1 target maps to the sum of those two transitions (imagine /ia/ and /au/ from Figure 7 overlain one on top of the other to create /iau/'s F1 target in Figure 8).

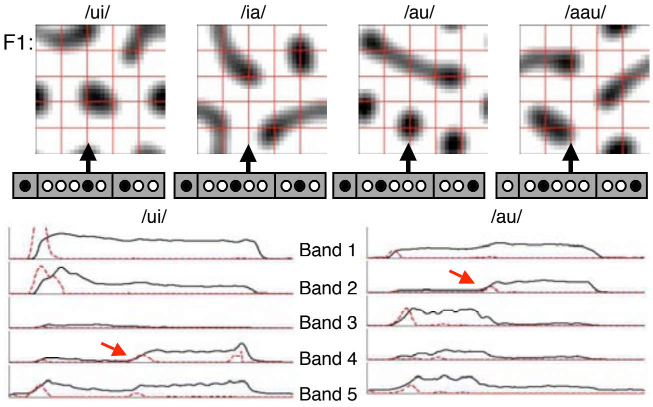

Figure 7: associating static input arrays with static F1 targets (top); frequency band plots for /ui/ and /au/ (bottom)

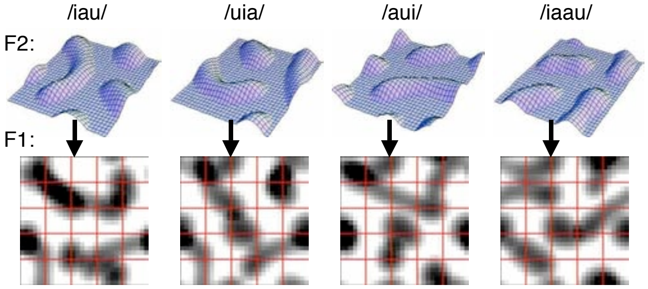

Figure 8: equilibrium states of F2 as control signals trained to prime corresponding F1 targets

Processing essentially works as follows: F1 goes to an equilibrium corresponding to an input transition and maintains this equilibrium until fatigue sets in. From fatigue, the equilibrium eventually spontaneously collapses. Under the influence of top-down priming from F2, F1 then immediately shifts toward a new attractor equilibrium that maps to the other transition that is 'primed' by top-down activation from F2. Sensory input also has an influence.

To evaluate training, the activations of motor category nodes were examined through time as the network responded to manual activations of F2. Specifically, the network including spectral model feedback was initialized to the values it had from a run at a snapshot in time just after the phrase "got to be" (recall Figure 6). From there, F2 was forced into its equilibrium corresponding to one of the test words. Figure 9 presents a summary depiction of the activations of the motor nodes in response to these four situations. Top-down influence of /iau/ activation in F2 (top left of figure) resulted in an on-off response of second motor category node followed by the third motor category node to theoretically generate the tongue movement sequence for /iau/. Likewise, the /uia/ signal from F2 produced motor node activations corresponding to the /ui/ and then /ia/ transitions (top right of figure). Results were successful for /iaau/ (bottom left) and /aui/ (bottom right) as well.

There are quite a number of issues to discuss. For instance, notice how the production of /au/ in /iau/ occurs earlier than the production of /au/ in /iaau/. This is due to the function of the spectral model. Top-down influence from the F2 pattern for /iaau/ was hampered until inhibition from the timing expectancy node subsided. For a more detailed description of this and other aspects of the model - including animations and sound files corresponding to this paper - please visit: http://www.fluidbase.com/mike/ART-STiM

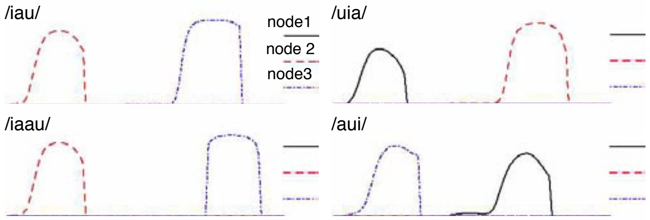

Figure 9: activations of motor category nodes through time as network responds to the four F2 feedback patterns

Conclusion

The model illustrates how a persistent control signal for a simple motor sequence of two articulatory gestures (a gesture as the transition between two vowels) may be implemented. As depicted in Figure 8, control signals for four gesture-related words are the equilibrium states of F2. By initializing the network from the same conditions but with these four different patterns of persistent F2 feedback, interactive dynamics result in the simulated production of the four different words. Note that these F2 equilibrium states do not lend themselves to phonological description. A linguist would be hard pressed to find patterns in those F2 states that could map to a phonological coding scheme.

A theoretically scaled version of the model allows for longer sequences of gestures and for increased phonetic complexity. Fields may be added in parallel to the network (i.e. as extensions of F1) to respond to change patterns associated with voice onset time, tongue flaps, labial movements, and other phonemic features. Fields may also be added in series (stacked on top of the network, i.e. adding an F3 and F4 etc.) to allow for longer streams - or sequences of sequences to form. Multiple processing streams may be realized by creating multiple stacks of fields and these streams may interact with each other through lateral connections. Numerous spectral timers may also be used.

This sketch of a scaled version of the model supports a construction grammar (Goldberg, 2003, Tomasello, 2003) approach to speech. In contrast to generative grammar, where a detailed motor plan is somehow assembled by means of a formal system, construction grammar views language production as "a repertoire of complex patterns that integrate form and meaning in conventionalized and often non-compositional ways." An utterance can be imagined as the result of how a variety of persistent activation patterns or exemplars at the top levels of different streams combine through network dynamics to produce motor output. Top-level fields correspond to concepts and grammatical regularities and act to contextualize each other. Phenomena such as over-generalization ("I goed to the store"), tongue twisters, and spoonerisms may better be appreciated from this integrated conceptualization.

As noted earlier, the Theory of Articulatory Phonology is not seen to be in conflict with the view taken in this paper. That is, if we return to consider speech communication and complex motor control in terms of units, we might now distinguish between *units of planning* versus *units of production*. Articulatory Phonology is essentially a framework for theorizing about production units or 'pre-coordinated action molecules.' Such a perspective is in full harmony with the use of motor category nodes in the model presented here. However, this harmony vanishes when planning units and production units are assumed to be isomorphic. This paper depicts the motor plan as a process rather than a product and as such the motor plan cannot be decomposed into units or analyzed using tree diagrams.

A new era of cognitive robotics is upon us. Because speech is ultimately a problem of motor planning and production, a fresh look at speech and language in terms of robotic control should provide new insights. With robots, the worldly interface cannot be assumed away. The modular approach of translating perception and action to and from the symbols of a formal system is attractive for a variety of reasons, but integrated control involving feedback loops and distributed processing is probably more in line with the way the brain works.

Acknowledgements

This work was partially supported by an NSF East Asia and Pacific Summer Institute (EAPSI) fellowship.

References

Beer, R. D., Chiel, H. J., Quinn, R. D., Espenschied, K. and Larsson, P. (1992). A distributed neural network architecture for hexapod robot locomotion. *Neural Computation, 4(3),* 356-365.

Brady, M. C. (2006). Adaptive resonance situated for speech learning and synthesis. *Proceedings of IEEE 9th International Conference on Development and Learning.*

Brady, M. C., and Port, R. F. (2007). Quantifying vowel onset periodicity in Japanese. *International Congress of Phonetic Sciences,* Saarbrucken. 337-342

Browman, C.P. and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica, 49 (3-4),* 155-180.

Goldberg, A. (2003). Constructions: a new theoretical approach to language. *Trends in Cog. Sci. (7),* 219-224.

Goldstein, L., Byrd, D., Saltzman, E. (2006). Vocal tract gestural action units. In Michael Arbib (Ed.) *Action to Language via the Mirror Neuron System.* Cambridge.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., Byrd, D. (2007). Gestural action units slip in speech production errors. *Cognition/Elsevier. 103 (3),* 386-412.

Hersch, M. and Billard, A. (2006). A biologically-inspired controller for reaching movements. *IEEE International Conference on Biomechatronics.*

Hockett, C. (1955). *A manual of phonology*. Chicago: University of Chicago.

Saltzman, E., & Kelso, J. A. S. (1987). Skilled actions: a task dynamic approach. *Psych. Review, 94,* 84-106.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition.* Harvard University Press.

[1] $\phi = 1 / \left(1 + e^{timespan - \sum_{t=0}^{-\infty} \sigma(u_t) - .5}\right)$

[2] $\sigma(x) = \frac{x}{x+1}$; $\sigma(x) = 0$ if x is negative

[3] $\lambda(d) = e^{\frac{-d^2}{r^2}} \cdot \left(\cos\left(\frac{\pi d}{2r}\right) - z\right) \cdot \left(\frac{1}{1-z}\right)$

[4] $h = -.27, \quad r = 4.5, \quad z = .15$

[5] See page 322 of PDP V1, Rumelhart & McClelland. (1986).