A Bidirectional Graphical Model for Babble-Feedback Learning in Speech

Michael Connolly Brady
University of Notre Dame
mbrady8@nd.edu

Abstract
A theory of motor-feedback learning based on a bidirectional-edge graph is summarized. The graph is inspired as a model of the cortico-thalamic circuit. Output from the network drives the motors of an articulatory speech synthesizer. The network is first exposed to an ongoing stream of speech sounds where neural fields tune themselves to respond to repetitive co-occurring patterns in the input (phonetic features). The network is then made to actuate motors with babble feedback learning in an effort to train the network to produce the speech sounds and strings of sounds that the network was originally exposed to. The model furthermore makes use of a bank of resonators, intended to depict the functional role of the cerebellum. This cerebellar model helps in coordinating the longer-term timing of network dynamics and helps to smoothly string together patterns of speech sounds through time.

Keywords: graphical models, reservoir computing, neural field model, articulatory speech synthesis, cerebellum.

1 Introduction

One popular approach to motor planning and production for speech relates to the theory of Articulatory Phonology (Brownman & Goldstein, 1995). Articulatory Phonology (AP) posits that a gestural score, akin to a multi-voice musical score or multi-track music recording, serves as a motor plan to be executed by a motor production model. Each track is taken to correspond to a separate component or dimension of the motor plan. Articulatory Phonology views phonetics as a high-dimensional description and phonology as a low-dimensional description of the same speech signal. This is in contrast to another popular view of speech motor control, where the control signal or motor plan is idealized as a sequence of holistic phonological units. These phonological units are translated into motor movements one-by-one as they arrive, like beads on a string. The Directions into Velocities of Articulators (DIVA) model of speech motor production (Tourville & Guenther, 2011) is in this camp of thinking and is a leading alternative to Articulatory Phonology. DIVA addresses three limitations of Articulatory Phonology. One is that DIVA incorporates sensory-motor feedback into its processing. The second is that DIVA is used as a model to guide brain imaging and functional neuroanatomy studies. The third is that DIVA depicts how speech is acquired in infancy. These two general views of the motor plan for speech find analogs in other domains of motor control.
The bidirectional-edge-graph approach summarized in this paper may be considered as a synthesis of the best parts of the DIVA and the AP models. Like DIVA, the approach here uses auditory training and motor babble feedback learning such that the system ‘learns to produce the speech sounds that it wants to hear itself say.’ However, rather than view the speech control signal as composed of holistic phonological units, the assumption taken here is more in line with Articulatory Phonology. That is, motor output for speech is dynamically generated as the control signal for speech emerges from multiple sources. This approach also applies to other domains of motor learning and control.

2 Modeling Overview

Figure 1 presents a graphical model of the cortico-thalamic system. Also included is a cerebellar model, where the functional role of the cerebellum is depicted using a reservoir of resonators. The model is embodied with an articulatory speech synthesizer. Neural fields are represented as vertices in the graph and sets of connections in white matter between neural fields are represented as edges. Regardless of where in the brain a piece of neocortex is located, that piece of neocortex is assumed to follow a standard pattern of functional circuitry (see: Mountcastle, 1957, Hawkins, 2004). It is thus justified and helpful to make the simplifying assumption that all vertices may be implemented based on a single neural field model. The neural field model has adjustable parameters to reflect regional differences in cortical function. What allows the abstract network of Figure 1 to learn to predict patterns is the automated adjustment of sets of weights or adaptive filters, represented in the figure as edges, or lines between vertices. Neural fields quickly fall into equilibrium states and are perturbed from those states based on input. Learning associates the equilibrium state of a field with its environment. Primary fields tune themselves to fall into systematic equilibrium states in response to systematic combinations of input across sensory input channels. Secondary neural fields then become able to tune themselves to respond to their environments once primary fields have settled into predictable behaviors. With experience, the network forms representations as each neural field systematically responds to its environment through time. Learning depends on connectivity and history of network exposure to input. It is important to note in Figure 1, that an edge is extraverted bidirectional; it carries two streams of information and in opposite directions. The dyadic nature of the edge is depicted using filled versus open arrowheads. One stream has a feedforward nature (referred to as driver signals, and depicted with open arrowheads), and the other has a feedback nature (referred to as modulator signals and depicted with filled arrowheads).

Figure 2 illustrates the basic mechanism of the network. In short, a field is a grid of processing units and each unit follows a dynamic field equation (Equation 1). Two neural fields interact with each other using driver \((d)\) and modulator \((m)\) signals. Like a Hopfield network, an isolated neural field with well-specified parameters quickly settles to a non-zero, non-saturation equilibrium state given random initial conditions. Input to the field may perturb the field to a new equilibrium state. This input comes from the combination of the driver and modular signals. As specified in Equation 2, the driver signal is based on the change in activation of its source units, passed through an adaptive filter. The modulator signal is found as the running average activation of its source units, passed through an adaptive filter. Thus, a driver signal hovers around zero when its source field is in an equilibrium state but this driver signal can spike to large values when its source field is perturbed. A modulator signal is consistently non-zero but will change to a new output pattern when its source field is perturbed.

An inspection of the modeling equations may help to clarify matters. Again, a neural field is implemented as a small sheet of cerebral cortex that is composed of a grid of discrete columns or units where each gridpoint or unit is updated per time step using Equation 1. This general equation and variations of it are widely used in dynamic systems and brain modeling, e.g.: Grossberg, (1976, 2003), Amari, (1977), Hopfield, (1982), Hinton & Sejnowski, (1986), Schöner & Kelso (1988), Beer, (2000), Kohonen (2000), Maass et. al., (2002), Jaeger, (2003).
A Neural Field Model for Babble-Feedback Learning in Speech

Michael Brady

Figure 1: a bidirectional edge graphical model. Each vertex is a dynamic neural field and each edge is a reciprocal connection in white matter. Output from the network drives an articulatory speech synthesizer. Sensory feedback is used to train the network. Timing feedback from a bank of resonators is included to depict the functional role of the cerebellum.

Figure 2: basic mechanism of the network. Two neural fields, F1 and F2, each in an equilibrium state, are poised to respond. A driver signal (d) that arrives to F1 may perturb F1 to a new state. If this happens, a driver signal from F1 is generated that may perturb F2. If F2 is perturbed, this will update the modulator signal (m) from F2 back to F1.

\[
\dot{u}_i = -u_i + S_i + h + n + \sum_j w_{ij} \cdot \sigma(u_j)
\]

(1) \quad \text{if } x < 0, \sigma(x) = 0, \text{ else } \sigma(x) = \frac{x}{x + 1}

\[
S_i = g_d d_i \left( g_a d_i + g_p m_i - (g_{shunt} \cdot \mu_i) \right)
\]

(2) \quad m_i = \sum_k w_{ik} \sigma_k

\quad d_i = \dot{d}_i = \sum_k w_{ik} \sigma_k

To put Equation 1 into words, the change in activation of a unit or column, \( u_i \) (of index \( i \)) at a given time step is determined as the sum of influence to the unit at that time step minus the activation of the unit from the previous time step. Influence to a unit at a time step comes from an input signal, \( S \), the field’s slightly negative bias, \( h \), a noise term \( n \), and from other units within the field. Influence from other units within the field is determined as the sum of the squashed activations of neighboring units (of index \( j \)) multiplied through corresponding within-field connection weights \( w \). A squashing function, \( \sigma() \), is used such that only units with non-negative activations can influence their neighborhoods. Within-field connection weights are specified as on-center off-surround by a Mexican hat weighting function. Input to the function is the Euclidian distance between two units and output of the function specifies their connection strength. Close neighbors excite each other, neighbors further away inhibit each other and neighbors further away still have no direct influence on each other. To allow for a modular approach and for the convenience of providing the same symmetric neighborhood of connectivity to all units of a field, an implemented field can be mathematically shaped like a torus. That is, an activation pattern extending off the left edge of a field can continue starting on the field’s right edge, and a pattern ending on a field’s top edge can continue starting on the field’s bottom edge. Level of activation of the cortical column is interpreted as the amplitude of gamma resonance of a
small vertical section of gray matter. A column is not assumed to have explicit boundaries in the continuity of cortex. Rather, a column may consist of thousands of interacting cells and portions of a column’s cells may be considered as belonging to other columns. With this in mind, a neural field is modeled using discrete columnar units with fine enough resolution that columnar continuity is approximated. Likewise, though time is continuous, if updated in discrete steps with fine enough resolution, temporal continuity is also approximated.

A neural field may be interpreted as something of a probabilistic switch. The field sits in an equilibrium state, waiting for a feature it has tuned itself to detect. Detection comes in the form of a driver signal that will perturb the field. That perturbation creates driver signals that may perturb other fields. Thus, the perturbation of one field may generate chains of perturbations in the network. These perturbations, or flips-of-switches, is a form of memory that alters the modulator environment of the network. Modulator signals influence or predict how driver signals will perturb fields in the future.

Equation 2 specifies how the driver signal, $d$, and the modulator signal, $m$, combine to provide input to the units of a field. Here, $\bar{u}$ is the running average activation of the unit being updated, and $g$ denotes signal-specific gain terms. Subtracting a portion of the running mean activation of a unit from the unit’s modulator input provides a shunting mechanism. Shunting specifies that the more active a unit has been, the less that unit will respond to continued input (or, conversely, the more suppressed a unit has been, the more eager it will be to respond). The modulator signal is found simply as the squashed running average activation of input nodes or zones, $o$, passed through a set of weights. The driver signal is found in much the same way, except here we care about the change in the signal rather than the running average of the signal. For simplicity of notation, a signal when italicized assumes the nature of change or running average in the signal’s origin field or fields.

Output from a field is based on the generalized activity of units within the field. Specifically, a field is sectioned into regions or zones where each zone’s activation is the average of the squashed activations of the units within it. A zone is composed of a small neighborhood of units. The average activation of a zone serves as the basis for input, $o$, to units in other fields (see Equation 2).

The network employs unsupervised learning to adapt to itself and to the world. Weights of filters are updated at each time step to associate the perturbations and equilibrium states of fields with one another. Equation 3 expresses this. A learning rate, $\eta$, scales the product of the sending zone value (a changing driver source or a running-average modulator source), the change in the receiving unit, the activation of the receiving unit, and a limiting term. The limiting term is based on the current weight such that as a weight approaches a maximum absolute value, the change in weight approaches zero. The change in activation of the target unit is included in the equation to specify that learning will occur mainly only during a target field’s perturbation. Modulator learning is geared at predicting and helping to realize a field perturbation based on the current running-average environment, while driver learning occurs when a target field’s perturbation occurs in synchrony with a non-zero driver signal (driver weights are updated when a source field is perturbed).

$$\Delta w_{ki} = \eta \cdot \dot{o}_{k} \cdot \dot{u}_{i} \cdot u_{i} \cdot e^{-g(w_{ki}^2)}$$

$$\Delta w_{ki} = \eta \cdot \dot{o}_{k} \cdot \dot{u}_{i} \cdot u_{i} \cdot e^{-g(w_{ki}^2)}$$

In pre-babble training, speech sounds are fed to the network. A training signal is derived from human speech audio recordings, pre-processed into a low-dimensional signal. This signal may involve the first few formants of the speech signal with F0 pitch and voicing information, Mel-scaled speech cepstral coefficients, or as 6-channel and 12-channel simulations of cochlear implants. During pre-babble training (no motor output), the network adapts its filters and essentially learns to ‘perceive’ patterned changes in input channels by simply being exposed to those changes.
After pre-babble training, motor output is generated by running unit activations of all fields through one large adaptive filter to an array of motor output nodes. Each motor node pair controls a degree of freedom of an articulatory speech synthesizer. This articulatory synthesizer is based on the Maeda synthesizer (Maeda, 1982). During a training scenario in babble learning, the values of the motor output nodes are pseudo-randomly set on movement trajectories while voicing is turned on. This generates a sound from the speech synthesizer. Auditory feedback from that sound (as well as proprioceptive feedback) propagates through the network to perturb fields. These changes in network equilibrium states are associated with changes in motor positions through the motor output filter such that future changes in these same network equilibrium states will produce these motor output trajectories. The idea is that the network learns to speak along with an input signal and eventually uses its own feedback to guide the articulatory synthesis. As will be discussed, training can be very slow due to extreme computational overhead. The training of a full-sized network can only be theoretical at this point based on today’s mainstream technologies. Current methods involve training a network in small parts, and working with minimal systems.

Sensory input is never to a quiescent system. A complex theoretical network of fields may form chains of perturbations and oscillating perturbation patterns in the graph, even in the absence of sensory input. These resonances are considered as the basis of observed alpha and beta rhythms in EEG recordings. One theory for the functional role of the thalamus is that it acts as a traffic controller to project and synchronize cortico-thalamic rhythms across the entirety of the cerebrum (see: Buzsaki, 2006, Sherman & Guillery, 2006). The timing of these rhythms is theorized to play an important role in predicting when a perturbation or cascade of perturbations will occur, in addition to what will occur.

The [lateral] cerebellum is modeled as a reservoir of adaptive resonators that helps to better coordinate cortico-thalamic rhythms across relatively long time scales. The cerebellum is somewhat interpreted as a sensory organ. Input to the organ is from the brain ‘proper,’ and output from the organ is a signal resulting from the conversion of spatial patterns of cerebral activation into predictive spatial-temporal patterns that influence cortico-thalamic processing in the same way that traditional sensory input does. The foundational mechanism of the cerebellum is considered as the circuit that involves the inferior olive, its corresponding climbing fibers, a Purkinje cell, and deep cerebellar nuclei, among other components. This circuit is modeled to resonate, and collections of these circuits are thought to synchronize with each other through gap junctions in the inferior olive (see Llinas, 2010). These circuits run on a relatively long time scale (with a period in the range of 250 milliseconds). Information from cerebral cortex arrives along parallel fibers in cerebellar cortex to influence which resonators or olivocerebellar circuits will become active to provide cerebellar feedback to the cerebrum, as relayed through the thalamus. It should be noted that there are a large number of theories relating to the functional role of the cerebellum. The view introduced here is that of this particular author. For a more detailed introduction to this model of the functional role of the cerebellum, a review of competing cerebellar theories, and a more detailed discussion and analysis of the general cortico-thalamic circuit described in this paper, see (Brady, 2009, 2012).

3 Discussion

The overarching aim of this work is to extend neural field modeling ideas into a bidirectional edge framework for speech motor planning and production. The bidirectional edge graph allows for conceiving global brain function in reference to a dynamic Bayesian networks perspective. Under this conceptualization, the vertex of the graph is not discrete, as in an HMM, nor is continuous, as in a Kalman filter, but is roughly categorical. A neural fields sits in an equilibrium state waiting to detect a pattern it has tuned itself to detect, this detection takes the form of a perturbation. The dyadic nature of the driver-modulator edge does not allow the graph to saturate with runaway feedback. Driver signals percolate information through the system while modulator signals simply act to influence the result of
driver signals. With this, timing information is encoded. A cerebellar model is also included in this global description of the theory. The output of the cerebellar model provides a predictive timing signal that helps to coordinate relatively long-time dynamics of the cortico-thalamic network. The cerebellar model becomes particularly useful when training a network on longer sequences of sound patterns.

It should be noted that the modeling approach summarized here is computationally intensive. As fields are added to a simple two-field network, the already intensive computing required increases exponentially. It is slow-going to develop and test even small networks, let alone medium sized ones and larger ones. The computational intensity of the approach is especially unattractive from a robotics perspective. Faster solutions are needed. Current work involves the development of a more computationally efficient method of approximating the neural field function. Another potential solution being explored is to implement the neural field model on GPUs rather than CPUs.

It is hoped that the general theory described in this paper provides some new insight as to how the motor plan for speech may be conceptualized from a biology-inspired modeling perspective.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27

Beer, R.D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*

Brady, M.C. (2009) Speech as a problem of motor control in robotics. *Proceedings of the Cognitive Science Society, Amsterdam*

Brady, M.C. (2012) *A Field-based Artificial Neural Network with Cerebellar Model for Complex Motor Sequence Learning*. (Ph.D. Dissertation) Indiana University

Browman, C.P. and Goldstein, L. (1995). Dynamics and articulatory phonology. In *Mind as Motion*, Port & Van Gelder (eds). MIT press

Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford, UK. Oxford University Press

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23

Grossberg, S. (2003). Adaptive Resonance Theory. *The Encyclopedia of Cognitive Science*. London : Macmillan Reference Ltd

Hawkins, J. (2004). *On Intelligence*. Times Books, Henry Holt & Co. New York

Hinton, G.E., & Sejnowski, T.J. (1986). Learning and Relearning in Boltzmann Machines, in D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Volume 1: Foundations (MIT Press)

Hopfield, J. (1982). Neural Networks and physical systems with emergent collective computational abilities, *Proc. of the National Academy of Sciences, vol. 79*

Jaeger, H. (2003). Adaptive nonlinear system identification with echo state networks. In *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, K. Obermayer (Eds). MIT Press, Cambridge

Kohonen, T. (2000). *Self Organizing Maps*, 3rd Edition. Springer

LLinas, R. (2010). Olivocerebellar System. In Shepherd, G. (Ed) *Handbook of Brain Microcircuits*, Oxford University Press

Maass, W., Natschlager, T., Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations, *Neural Computing*

Mountcastle, V.B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiology*. 20 (4)

Schöner, G., & Kelso, S. (1988). Dynamic pattern generation in behavioral and neural systems. *Science*

Sherman and Guillery (2006). *Exploring the Thalamus*. MIT Press, Cambridge, MA.

Tourville, J. & Guenther, F. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*