Adaptive Resonance Situated for Articulatory Speech Learning and Synthesis

Michael Brady
Cognitive Science Program, Indiana University
1910 E. 10th St., Eigenmann 819
Bloomington, IN 47406-7512, USA
mcbrady@indiana.edu

***Abstract* – A control framework for a robotic vocal tract is introduced. I incorporate a timing orchestration mechanism into a basic ART network and the network is situated within a speech perception-production loop. The network is first trained to auto-associate its motor output with its resulting sound input during a 'babble learning' phase. Further training involves alternating between babble learning and learning for a set of pre-synthesized rhythmic vowel patterns. Preliminary analysis indicates that the network is successful at distinguishing and reproducing the training patterns with their correct temporal structures.**

***Index Terms* – adaptive resonance theory, ART, speech processing, robot dynamics, cerebellar model arithmetic computers.**

I. Introduction

Adaptive Resonance Theory (ART), first introduced by Steven Grossberg in the mid 1970s [1-3], provides a strategic platform from which to model the learning and recognition and production of temporal patterns. The theory provides a biologically plausible account of the general workings of the brain and as a result, it has enjoyed a longevity in relevance. In keeping with its biological underpinnings of ART, I introduce a mechanism for orchestrating the combinatorics of neural firing patterns across time. Such a mechanism is needed in addressing how the events of temporal patterns like speech and music are to be integrated in the working memory or resonance of an ART network.

I situate the network in a speech perception-production loop. The architecture reflects how language acquisition may be interpreted as a problem posed to the field of robotics. Here, speech is considered as coordinated sequences of vocal articulator motor movements where symbolic notation systems are never needed or assumed. Raw sound is input from the environment and articulator commands are issued to the environment. Like musical riffs played on a saxophone or guitar, speech involves learned patterns of motor coordination with feedback. Taken as a whole, this paper introduces a modeling approach that embraces the popular anti-symbolic view of modern robotics as applied to speech.

I begin with a brief synopsis of adaptive resonance. ART is based on collections of cells, called *processing units*. Processing units are grouped into *fields*. How units interact with each other within a field and across fields is the basis for the emergent behavior of the system. Fig. 1 depicts a simple ART network where circles represent processing units and rectangles around them represent fields, labeled F1 and F2. An orienting subsystem is also depicted.

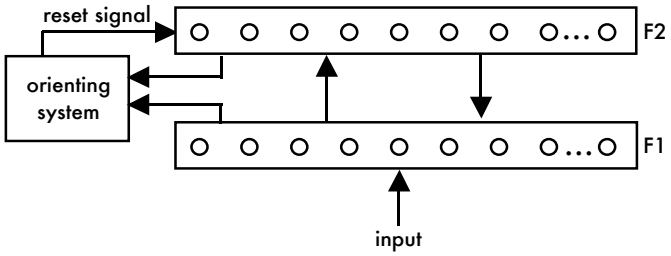

Fig. 1. Basic ART network

Though processing is conceptualized to happen continuously, the network is updated in discrete time steps or at each tick of a clock. During an update, input to F1 from the outside world mixes with top-down input from F2 so that new activations for processing units in F1 are determined. Input to a processing unit in F2 for that same time step is calculated in the standard connectionist way: as a function of the vector of F1 activations multiplied and summed through sets of weights corresponding to F2 units. Separate sets of weights fully connect units in F2 back to units in F1. The sets of weights between fields are referred to as *adaptive filters* and are indicated with arrows between the two fields of Fig 1. Weights of the adaptive filters are adjusted through a form of Hebbian learning. A pattern of activation in one field becomes associated with a corresponding pattern of activation in the other field. Input to a field from the other field (and from the outside world in the case of F1) is one of the two main forces that determines a field's pattern of activation. The second force has to do with how units within a field interact.

Units of a field compete with each other for activation through on-center, off-surround connectivity. An active unit will inhibit its neighbors in the field such that a unit that begins with a slightly higher activation will achieve 'winner-take-all' status by successfully suppressing all of its neighbors and thus reducing the amount of inhibition it receives. Before training, weights are initialized with random values. Some pattern of activation in F1 in this condition will provide input to all processing units in F2 where out of blind luck some of the F2 processing units will receive slightly higher input values than others. In this way, a few units will become very active while the rest are suppressed to become inactive. Winner-take-all interaction between units in F2 then 'contrast enhances' the activations that resulted from input from F1.

Because of the nature of the Hebbian algorithms used, learning between fields is essentially only enacted on weights when receiving units are active; adaptation only occurs for weights projecting to units that have won the competition for activation. In this way only a relatively small number of weights between fields are updated in associating activation patterns between fields. Furthermore, learned patterns of activation become *attractors* in the sense that a random pattern of activation will drift into a familiar pattern of activation within a field. With training, a pattern of activation of F1 processing units gives rise to an associated pattern of contrast-enhanced activation of F2 units, which will in turn influence how F1 responds to input at the next time step. This 'resonance' that emerges between F1 and F2 guides the processing of input as it arrives from the outside world. Resonance between fields is the short-term or working memory of the system and is argued by Grossberg and others to be the basic mechanism of consciousness.

If resonance translates to consciousness, the orienting system translates to attention. The job of the orienting system is to detect a gross mismatch between F1 and F2. If the pattern of activation in F2 no longer corresponds to input from the outside world, F1 will be receiving divergent messages from its two input sources. In this case, the orienting system kicks in to wipe out activations in F2. This gives suppressed units in F2 the chance to become active in response to the new activation patterns arriving from F1.

For a much more complete overview of ART, the reader is directed to [3]. For an introduction to ART for temporal pattern processing, I recommend Robert Gjerdingen's ART por l'Art model [4-5]. His implementation learns to develop expectations for Mozart Melodies.

II. ARESA

To think about ART in terms of brain anatomy, consider F1 to map to the thalamus and brainstem while F2 maps to cerebral cortex. The orienting system then is taken as the hippocampal formation and basal forebrain. Though this may be a simplistic description, it helps in talking about ART in anatomical terms. I next place the entire system into a perception-production feedback loop for speech and I describe a timing mechanism I have developed that maps to the cerebellum in this anatomical framework.

The model described in the last section is the foundation for ARESA (Adaptive Resonance Embodied for Speech Acquisition). Fig. 2 depicts ARESA as the same basic ART network of Fig. 1 with the innovations. F1 is now fully connected through a set of weights to a set of output or synthesis parameter units. The pattern of activation in F1 thus informs the controllers of a robotic vocal tract[1]. These parameters correspond to: 1) air pressure 2) vocal cord tension 3) tongue front/back 4) tongue high/low and 5) jaw position. The front/back high/low tongue positions are determined by

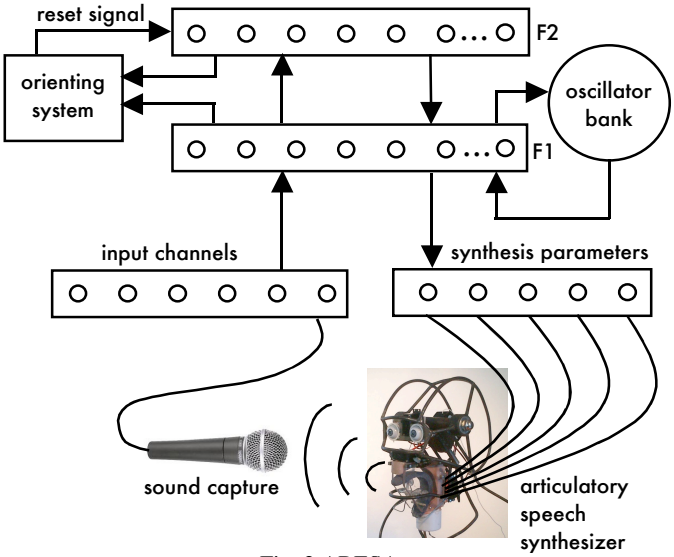

Fig. 2 ARESA

two motors that move a silicon tongue up and down and/or forward and back to change the effective shape of a tube. A third motor controls the height of the jaw and the height of the jaw in turn influences tongue position within the tube. An electric speaker mounted on the bottom of the tube simulates vocalization by making a buzzing sound within a frequency range. The frequency of the buzzing more or less corresponds to vocal cord tension and amplitude of the buzz more or less corresponds to air pressure. Interaction between air pressure and vocal cord tension also influences the spectral qualities of the buzzing sound. Sound output from the synthesizer is fed through a bank of bandpass filters to supply input back to F1. The input channel mechanism can be thought of in terms of a cochlear implant. Modern cochlear implants typically have from 6 to 16 input channels and people with cochlear implants can generally distinguish the sounds considered in this paper. The tract with filter bank forms a perception-production feedback loop where patterns of activation in F1 relate to both input and output. Though the model is illustrated and motivated to drive a robotic vocal tract, the robotic tract currently serves mainly as a poster child for the project. The work reported in this paper is based on simulations using a Klatt digital speech synthesizer with roughly the same control parameters.

Training starts by simply having the system babble. All weights and unit activations and synthesis parameters of the system start out with random values. With the synthesis parameters randomized, the synthesizer then generates a sound based on random articulator positions. The sound is band-pass filtered into six channels (each channel the output of a third order Chebyshev type II filter). Values from the channels at that time step are used as sending-unit activations to input F1. Sending units are fully connected through a set of weights to the units of F1. A form of Hebbian learning is used to associate the input pattern with the still random pattern of F1 unit activations. This learning follows (1) where the change in weight between two units, Δw_{ij}, is found as a product of the learning rate, η, the activation of the sending unit, a_i, and the

[1] Please find video clips and a description of the mechanics of the robotic tract and latest model developments: http://www.fluidbase.com/ARESA

difference between the activation of the sending unit and the connection weight, $(o_i - w_{ij})$:

$$\Delta w_{ij} = \eta a_j (o_i - w_{ij}) \quad (1)$$

Channel values or activations are then passed through these updated weights to provide input to F1 with (2). Activation, a, of a receiving unit, j, is calculated as a function of the sum of sending unit activations, i, times their corresponding weights, w.

$$a_j = f \sum_i a_i w_{ij} \quad (2)$$

Raw input to a unit is squashed with a squashing function, f, from a value between 0 to ∞ to a value between 0 and 1.

The F1-to-F2 and F2-to-F1adaptive filters and F1 and F2 unit activations are then updated as in Gjerdingen's model [6]. With feedback from F2, a new pattern of activation in F1 is determined. This pattern is used to update the weights from F1 to the synthesizer based on synthesizer parameter values from the beginning of the time step. The parameter values are then each slightly shifted in random directions and a new iteration of learning begins for the next time step. The process continues until there is no significant weight change from time step to time step. After 'babble training', a pattern of activation in F1 corresponds to both perception and production. When ARESA hears a static vowel sound, it simultaneously knows how to produce it.

Oscillator Bank

A major innovation to the earlier ART model is now noted. A temporal orchestration mechanism is the timekeeper of the system. The mechanism is based on a bank of adaptive oscillators where each oscillator has a distinct natural period it wants to oscillate with. This period, ρ_i, is assigned as the index of the oscillator multiplied by the 'spread rate' or difference between oscillator periods, β:

$$\rho_i = i\beta \quad (3)$$

The oscillator bank is conceptualized to represent a population of circuits where the preferred natural period of the bank is distributed around a 'tactus' periodicity, or a periodicity of about twice per second. Oscillators with natural periods closer to a specified tactus period will have a stronger role in perceiving a pattern while more irrelevant oscillators, with natural periods of say, close to nothing or of over a full second will have almost no influence. The population strength, λ, of each oscillator in the distribution is given with:

$$\lambda_i = e^{\frac{-(\beta i - \tau)^2}{\beta^2 \sigma}} \quad (4)$$

where τ is the period of the mean tactus and σ is a constant determining the distribution's shape.

An oscillator is described as a pendulum. As an oscillator travels around a phase circle, its phase increases from 0 to 1 (in radians). When it reaches a phase of 1, it is considered to be back at a phase of 0 again. At each time step, the mean input to F1 is calculated and this provides a single input value to the oscillator bank. For each oscillator, this input directly increases or decreases the oscillator's current amplitude, α_t, from the amplitude the oscillator had at the previous time step, α_{t-1}, based on its previous phase, θ_{t-1}:

$$\alpha_t = \sqrt{(I + \alpha_{t-1} \sin(2\pi\theta_{t-1}))^2 + (\alpha_{t-1} \cos(2\pi\theta_{t-1}))^2} \quad (5)$$

Because input increases only the 'momentum' of an oscillator, a second equation is necessary to shift the phase to compensate for the shift in amplitude such that the motion of the oscillator remains consistent from time step to time step:

$$\theta_t = ar\cos(\frac{\alpha_t \cos(2\pi\theta_{t-1})}{\alpha_{t-1}})(\frac{1}{2\pi}) \quad (6)$$

It is also necessary to keep the amplitude of the oscillators from growing without bound. The updated amplitude of an oscillator, α_u, is kept in the range of 0.0 to 1.0 with the following squashing function, where κ is a constant:

$$\alpha_u = \frac{\alpha\kappa^\alpha}{\alpha\kappa^\alpha + 1} \quad (7)$$

After input to an oscillator is calculated and squashed, the amplitude decays:

$$\alpha_u = \alpha - \frac{\alpha D\beta}{\rho} \quad (8)$$

where D is a global decay rate. Dividing by the period of the oscillator standardizes the decay for all oscillators, as oscillators with longer periods are updated more often in relation to their cycles.

Finally, at each time step the phase of an oscillator must advance a portion of its period, where T is the duration in milliseconds of the time step (simulations reported in this paper were run in time steps corresponding to 5 ms)

$$\theta_{t+1} = \theta_t + \frac{T}{\rho} \quad (9)$$

Output from the oscillator bank at each time step is a single value that represents the overall state of the bank at that time step. The optimal time for an oscillator to receive its input such that amplitude is maximally increased is at a phase of what I call its 'firing phase' (firing phase = 0.25 in radians, based on Eqs. 5 & 6). Each oscillator that passes its firing phase during a time step contributes to the output of the

oscillator bank based on its amplitude and its relevance in the distribution. That is, the output value of the oscillator bank, O_B, at a time step is found as the sum of each oscillator's firing, F (if an oscillator of index i passes its firing phase during a time step, F_i=1.0, else F_i=0.0), multiplied by that oscillator's amplitude and population strength:

$$O_B = \sum_i F_i \lambda_i \alpha_i \tag{10}$$

O_B is squashed with (7) and then serves in a threshold function for the output cell of each F1 unit. More specifically, input to each unit of F2 by eq. (2) becomes something like:

$$I_j = f \sum_i O_B a_i w_{ij} \tag{11}$$

This enacts what may be viewed as a periodic thresholding of F1 when there is periodicity in the input so that sound segments are grouped or fused together. Information is released from F1 to F2 as it occurs within 'perceptual pulses' of the oscillator bank. Events leading up to the pulse are commingled into a single firing pattern. I return to this in the discussion section.

The important point to note here about the oscillator bank is that it falls into specific behavior patterns given specific temporal structures of input patterns. When periodicity exists in an input, that periodicity is captured by the output. If there is no apparent periodicity in the input, but the input still exhibits rhythmic regularity, that regularity is also distinguished by the system.

III. Initial Results

A bank of twelve test patterns were pre-synthesized by the simulated robotic vocal tract and used to train the system. The training patterns were made of three basic vowels (/a/, /u/, and /i/) presented in two different rhythmic contexts. Table I presents the twelve possible permutations of rhythmic-orderings of the vowels. There is an isolated vowel and a vowel pairing in each pattern. Dots between vowels represent the analog of musical rests where the vowel sounds and rests are all of equal duration. Including the rests as events, there are six events of equal duration in each pattern. The difference between the two rhythms is that one rhythm has the long rest after the isolated event while the other rhythm has the long rest after the event pair. There are two possible sequential orderings the three vowels could take (i.e. either the /u/ or the /i/ can follow the /a/) and each of the vowels could be the isolated event. The patterns were each repeated three times to form the stimuli for the test.

For the sake of this simulation, twelve supervised training nodes were added to the system in the form of category nodes on F2. Each of these nodes corresponded to one of training pattern of Table I. When a specific training pattern was presented as input, that pattern's training node was also turned on. Weights from F2 to the training node were adapted according to the generalized delta rule [6]. Training alternated between this supervised learning and babble learning. After training, turning on the training node and activating F2 with that training node's weights and waiting for resonance would generate sound from the synthesizer. Listening to these sounds provides a rough way to evaluate how the network has learned. The network is seen to easily reproduce the training patterns (as judged by this author).

TABLE I
TWELVE REPEATING VOWEL-RHYTHM INPUT PATTERNS

a..ui.	a..iu.	a.ui..	a.ui..
u..ia.	u..ai.	a.ui..	a.ui..
i..au.	i..ua.	a.ui..	a.ui..

A series of perceptual studies are underway. For example, in one experimental design, subjects are presented with pairs of stimuli (e.g. from Table I) and on a scale from 1 to 10 they are asked to say how similar the patterns are. Collective results provide a basis for evaluating the model. A similarity measure from ARESA is taken as a function of the amount of time it takes for the system to switch between resonances when exposed to one pattern directly after the other. The measure is highly dependent on the parameters and functionality of the orienting system. Preliminary results indicate that similarity ratings by an optimized model may resemble ratings made by people. However, a detailed discussion relies on a detailed discussion of the orienting system and is beyond the scope of this paper.

IV. Discussion

This paper is meant as an introductory overview to a large project. My intention is to provide a foundation for dialog on a number of issues relevant to the continued development of ARESA and on general topics in speech acquisition research. Future work involves continuing to refine and focus on various details and aspects of ARESA. These include 1) integration of events of temporal patterns with each other in working memory 2) issues involving attention and the orienting system 3) methodologies for evaluating the model in terms of human performance 4) the use of evolutionary algorithms to refine the architecture and learning parameters. I now briefly address these topics and attempt to tie everything together into a unified whole.

ART networks are inherently suited for the processing of temporal patterns. Yet, they have only ever really been implemented to process static and sequential patterns. More explicitly: I do not know of any ART network that can distinguish say, a sequence of musical notes presented in one rhythmic arrangement from those same notes presented in the same sequential order with a different rhythmic arrangement without a hack to encode temporal information in the symbolic representations of the notes (e.g. [4-5]). By incorporating a bank of adaptive oscillators that collectively generates expectations for the onsets of the high energy

portions of the speech signal (vowels), temporal relationships are mapped to an internal clock. Activations arriving from the input bank of bandpass filters collect into F1 and when the beat-tracker passes its firing phase, it raises activations above threshold. This is meant to mimic a kind of neural phase-locking. The generalization is that *trajectories* of speech sounds corresponding to articulatory gestures are gathered and released into F2 in cohesive batches. Input (and output) of ARESA may thus be interpreted as speech gestures or trajectories and are centered around transitions into (and anticipations of) vowels. Though a full discussion is beyond this introductory treatment, the phase-locking I introduce next is seen as orchestrated by neural circuitry associated with the cerebellum and perhaps a feedback loop of the cerebellum with the basal ganglia. See [7] for a relevant discussion on the role of the cerebellum in speech and cognition.

A temporal form of Hebbian learning serves as the foundation for conceptualizing neural phase-locking. Equation 1 introduced a form of Hebbian learning typically used in ART networks where the *strength* of a connection is updated based on activations of sending and receiving units. In what I term 'DHL' (dynamic Hebbian learning), the *timing* of a synaptic connection becomes the true mechanism of learning. Imagine an array of units connecting through synapses to a target unit. When connecting synapses fire and then the target cell fires, the connecting synapses that fired a little too early or a little too late should correct themselves. That is, synapses are modeled to adapt to fire a little more quickly or a little more slowly the next time they fire so as to optimize the recovery of neurotransmitters in reuptake. Through DHL, synaptic connections that are associated with each other become synchronized to fire at the same times. This increases the odds that the target cells associated with the synapses will receive enough activation to rise above their thresholds to fire. Resonances between fields in an ART network may thus be viewed in terms of phase-locking and the oscillator bank then can be interpreted as a preliminary mechanism to orchestrate these neural synchronicities. Work related to these ideas includes [8-9].

The orienting system is the least well developed aspect of ARESA. In ART, the orienting system maps to the role of the hippocampus and basal forebrain – it oversees resonance between the fields and thus it directs attention and learning. I anticipate the vast bulk of future work on ARESA will be geared at integrating the oscillator bank with the orienting system so that only strategic portions of F2 are reset upon F1-F2 mismatch.

The last development direction I pursue has to do with the use of genetic algorithms. Like the CTRNNs investigated by Randall Beer and others [10-11], the architecture and parameters and initial connections of ARESA may be optimized by generating, evaluating, and evolving network configurations. I am specifically interested in how weights between the filter bank and F1, and between F1 and the speech synthesizer may be hard-wired. This hard-wiring would vastly reduce the time needed for the system to learn and conceptually relates to the nativist view of language.

Situating an adaptive resonance network in a speech perception-production loop helps to illustrate some conceptual issues in current debate. For example, a motor theory of speech perception proposed by Alvin Liberman and colleagues [12-14] posits that intended articulatory gestures by a speaker are normalized by a listener based on neural circuitry that is intimately linked with speech production circuitry. In short, the theory proposes that we recognize the words of an utterance regardless of who utters them (different people exhibit different acoustic production characteristics), because we perceive the acoustic stream in terms of how we would articulate it ourselves. Patterns of activation in F1 and resonance in the model simultaneously correspond to input and output and speech is perceived in terms of production knowledge.

ARESA also relates to the work of Frank Guenther and colleagues [15] where they have developed a model termed DIVA (Directions Into Velocities of Articulators), that develops knowledge about vocal production by listening to itself babble. Unlike the simplistic steady-state vowels used in this study, DIVA is geared at articulator trajectories. Guenther makes a case for acoustic rather than articulatory goals in vocal production. With ARESA however, there does not seem to be a clear distinction between articulatory and acoustic goals. Motor equivalence relations naturally fall out of the system - though direct motor feedback to F1 from the robotic vocal tract has yet to be implemented and tested.

It is clear that there are many details to be addressed. However, as a painter who is trained to cover the entire canvas with paint before focusing on the particulars, I have described the big picture of ARESA. Details may now be fleshed in with reference to a grand scheme and as details take on their precise forms, the big picture will continue to emerge and to influence the roles of other details still. The focus and masse process now begins.

REFERENCES

[1] Grossberg, S. (1976). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. Blgcl. Cybrntcs., 23, 121-134.

[2] Grossberg, S. (2003a). Resonant neural dynamics of speech perception. *Journal of Phonetics*, 31, 423-445.

[3] Grossberg, S. (2003). Adaptive resonance theory. The encyclopedia of cognitive science . London : Macmillan Reference Ltd.

[4] Gjerdingen, R. (1990) Categorization of musical patterns by self-organizing neuronlike networks. Music Perception, 7(4):339-370.

[5] Gjerdingen, R. (1992) Learning syntactically significant temporal patterns of chords: A masking field embedded in an ART3 architecture. Neural Networks, 5(4):551-564

[6] Rumelhart, D. E. and J. L. McClelland (1986). Parallel distributed processing: exploration in the microstructure of cognition.Vol. 1 & 2. MIT Press.

[7] Keele, S.W., Ivry, R., Mayr, U., Hazeltine, E., & Heuer, H. (2003). The cognitive and neural architecture of sequence representation. Psychological Review, 110, 316-339.

[8] Wang D.L., Freeman W.J., Kozma R., Lozowski A., and Minai A. (2004): Guest Editorial for Special Issue on Temporal Coding for Neural Information Processing. IEEE Transactions on Neural Networks, vol. 15, pp. 953-956.

[9] Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory and serial order. Psychological Review, 107, 127-181.

[10] Beer, R. (1995) On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior* 3(4):469-509. University Press.
[11] Beer, R. (2005). Parameter space structure of continuous-time recurrent neural networks. Submitted.
[12] Liberman, A. (1957). "Some results of research on speech perception." J. Acoustical Society of America 29 (1).
[13] Liberman, A. (1996). Speech: A Special Code. Cambridge, MA, MIT Press.
[14] Liberman, AM & IG Mattingly (1989) A Specialization for Speech Perception. Science, 243, 489--494.
[15] Guenther, F.H., Hampson, M., and Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. Psychological Review, 105, pp. 611-633.