Neural Circuits for Any-Time Phrase Recognition with Applications in Cognitive Models and Human-Robot Interaction

Richard Veale[1] **and Matthias Scheutz**[2]
riveale@indiana.edu mscheutz@cs.tufts.edu
[1]Cognitive Science Program, Indiana University [2]Department of Computer Science, Tufts University

Abstract

Humans are remarkably good at recognizing spoken language, even in very noisy environments. Yet, artificial speech recognizers do not reach human level performance, nor do they typically even attempt to model human speech processing. In this paper, we introduce a biologically plausible neural model of real-time spoken phrase recognition which shows how the time-varying spiking activity of neurons can be integrated into word tokens. We present a proof-of-concept implementation of the model, which shows promise both in terms of recognition accuracy as well as recognition speed. The model is also pragmatically useful to cognitive modelers who require robust any-time speech recognition for their models such as real-time models of human-robot interaction. We thus also present such an example of embedding our model in a larger cognitive model, along with offline analysis of its performance on a speech corpus.

Keywords: Liquid State Machine; Neural Network Model; Any-Time Speech Recognition

Introduction

The mechanisms that convert physical signals such as light and sound into firing rates of neurons are well-studied. However, the way these signals subsequently influence behavior, especially cognitive behavior, is less well understood. In particular, the progression from continuous real-time input streams at the physical transducer level to high-level cognitive processes that abstract over many physical characteristics is challenging, so much so that classical cognitive architectures simply assume higher-level representations such as word tokens instead of modeling the processes that generate them. While these assumptions do not pose problems for disembodied models (e.g., of higher-level cognitive processes such as analogical reasoning in language), they are critical showstoppers for embodied situated models that depend on being implemented on robots that interact with their environments in real-time (e.g., in the context of human-robot interactions in natural language). In such models, sensory processing must be performed one way or another, and while it is sometimes possible to substitute engineering solutions for biologically plausible sensory modules (e.g., artificial speech recognizers instead of biologically plausible models of speech recognition), those substitutions often come at a price that assumptions have to be made about the nature of the interface between those parts of the model that are meant to be biologically plausible and those parts that function as proxies for yet-to-be-developed biologically plausible parts. Specifically, the sensory modules must be able perform their function (e.g., recognizing words or objects) at *least* as well as humans, or else other components of the model have to account for errors in perceptual processing (e.g., word recognition errors). Moreover, the sensory module must be able to perform its task and make its result available *at least as fast as* the corresponding module would in humans to be able to respect human timing (e.g., the human expectation to hear a verbal acknowledgment at the right time in response to an utterance). As a result, these two requirements often pre-empt the use of traditional computational methods that perform sensory-input-to-token conversion.

In this paper, we will address the problem of biologically plausible real-time sensory processing of speech signals with a two-fold goal: (1) to provide a biologically plausible neural model of human spoken phrase recognition, and (2) to provide a sensory module that can be embedded in classical cognitive architectures for the study of embodied situated models of natural language interactions. Specifically, we propose a new approach to robust speech recognition which gives continuous access to meaningful partial results, and returns word or phrase tokens that can be directly used by higher-level cognitive models. The model includes several parts of the early auditory processing system in humans: the cochlea (converting time to frequency domain), parts of the olivary complex (applying several filters to the cochlear-processed signal), and a sensory cortical area (comprising a recurrent spiking neural circuit to integrate the signal). Category separation is performed by "readout neurons" (one per category) that respond continuously to ongoing activity in the recurrent circuit based on the particular weighted projection they receive, as is customary for the employed *liquid-state machine* (LSM) neural model (Maass, Natschlager, & Markram, 2002). The weighted projection received by each readout neuron is determined before-hand based on offline training on a speech corpus. The main contributions are thus the implementation of the speech processing neural architecture and the method of converting the instantaneous output of readout neurons to the token-type output for use by subsequent cognitive processes. The paper starts by laying out the background and the problems to be solved by a model that has to convert time-varying instantaneous neural-readout behavior to discrete categories. The subsequent model section then describes the neural speech-encoding and integration parts of the model. Then an analysis of the performance of the model on subsets of a speech-corpus from human-human interaction experiments is presented together with links to videos in which the model is used as part of a larger cognitive model for situated embodied human-robot interaction experiments where speech utterances are used to control the robot's ongoing behavior.

Background

Liquid state machines have been proposed as neurologically plausible models of cortical microcircuits (Maass et al., 2002) which can be used for time-invariant categorization of continuous-time signals by way of simple linear "readout units". The conversion from instantaneous readout activity to discrete tokens that accurately encode the temporally-extended category, however, is not a trivial task as the instantaneous activity of a readout neuron only reflects that the very recent activity (e.g., tens of milliseconds) of the recurrent circuit is similar to what its activity was *at some point* during its response to the category on which the readout neuron was trained. In other words, if a readout neuron R is trained to respond to an isolated category, e.g., a word W, and is firing vigorously to a stimulus S presented to the circuit, it is not clear which part of S is recognized by the readout as being similar to W. It is possible that the readout will fire (entirely by coincidence) in response to a 20 ms segment of S but remain silent otherwise (as that part of the speech signal bears resemblance to patterns that occur in typical speech signals for W, e.g., similar phones). Obviously, it would be premature in this case to conclude that S is an instance of the category W – after all, W might be typically 500 ms long, and parts of it will be similar to parts of many other words. This dissociation between recognizing matching parts of a word versus recognizing the whole word (or phrase, for that matter), is the first problem to be solved: the challenge here is to produce a mechanism for filtering out coincidental noise and choosing a single "winning" category from among all readout neurons that will have time-variant activations throughout the presentation of S.

The second problem to be solved is to determine *when* to select a winner (clearly a winner cannot be select at very small time intervals as this would be tantamount to recognizing new words all the time). Since a single stimulus can span thousands of time-steps at which neurons can fire, the model should return a single token exactly once per stimulus and only if the stimulus is one of the categories that it has been trained on. Even if there is a stimulus-length patch of noise, the model should not recognize it as being similar to the readout with the highest activation, but should not detect any word token at all (alternatively, it could detect a "noise token").

Liquid state machine models have been previously applied to cases where the speech corpus was pre-processed into frequency channels that were guaranteed to have only one spike per word (coding onset, offset, or peak of that frequency band) (Maass et al., 2002). However, these assumptions are unrealistic and it is difficult if not impossible to produce this type of encoding in real-time from raw audio streams. Another approach addressed this encoding limitation and compared the performance of LSMs using different sound-coding front-ends (Lyon cochlear vs. MFCC) as well as different methods for converting the front-ends' analog output into input spike trains for the LSM (Verstraeten, Schrauwen, & Stroobandt, 2005). While both approaches performed category-token recognition well by their own metrics (i.e., the ratio of the number of correct readout spikes to total time points in (Maass et al., 2002) and the class with the most readout spikes in response to a given word file in (Verstraeten et al., 2005)), neither method is applicable to real-time speech recognition, since real continuous-time audio is not separated into "files" with a clear stimulus onset and offset. Rather, it is non-trivial to detect the onsets and offsets of real utterances from continuous speech streams, which are often full of non-word noises, variations in word pronunciation, or words on which the system has not been trained.

Hence, we developed novel and realistic neural implementations of onset/offset detectors to increase recognition performance and aid in utterance detection and classification. Specifically, the model is based on the approach of Smith and Faser (2004) who, inspired by Ghitza (1987), present a biologically inspired onset-detection regime using depressing synapses. This implementation via short-term plasticity (STP) synapses is justified based on the evidence provided by MacLeod, Horiuchi, and Carr (2007), who argue that not only synaptic depression, but also facilitation, can play a critical role in auditory processing. Finally, for auditory input signal processing we utilize the Lyon cochlear model (Lyon, 1982; Slaney, 1998) which effectively applies band-pass filters and transformations to sound waves to approximate the firing activity of a set of neural channels along the cochlea.

Model Architecture and Implementation

Figure 1 is a visualization of the neural circuits implemented for speech recognition without the learned components (i.e., readout neurons and readout integrators). Here, we describe the components of the model together with all parameters used for the empirical evaluation presented later.

Neural and Synaptic Models

The neural model uses Leaky Integrate-and-Fire (LIF) neurons, whose membrane potential V_m:

$$\frac{\partial V_m}{\partial t} = \frac{-(V_m - V_{rest}) + R_m \cdot (I_{bg} + I_{syn})}{\tau_m} \quad (1)$$

where R_m is the membrane resistance, I_{bg} the background current, and I_{syn} the total current impinging from afferent synapses. $-V_m$ represents the leakage term, causing the membrane potential to decay exponentially with time constant τ_m. When V_m reaches the threshold value V_{thresh}, the neuron "fires" and V_m is reset to V_{reset} and enters a refractory period during which it does not update.

The model uses static or dynamic synapses to connect neurons. A static synapse has a post-synaptic response (PSR) that decays exponentially with time constant τ_{psr}. The dynamics of the post-synaptic response q_{psr} of a synapse is thus:

$$\frac{\partial q_{psr}}{\partial t} = \frac{-q_{psr}}{\tau_{syn}} \quad (2)$$

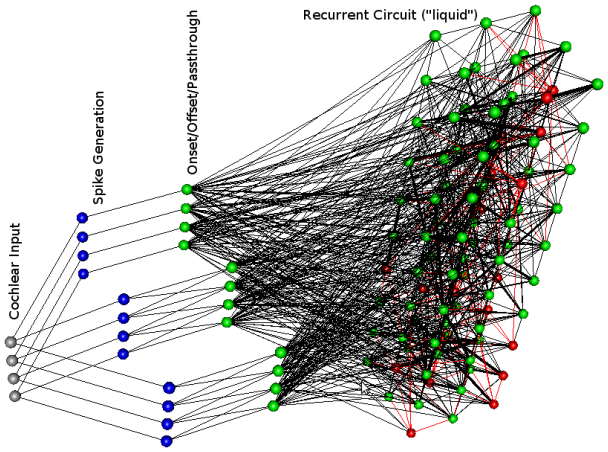


Figure 1: 3-D visualization the neural model described in this paper. The pictured circuit has only 4 input channels, and a $3 \times 3 \times 10$ recurrent circuit. The actual circuit has 52 input channels and a $5 \times 5 \times 15$ recurrent circuit. Readout neurons not shown (they would be on the right of the recurrent circuit, receiving input from it).

Synaptic dynamics (short-term plasticity, STP) are implemented following (Legenstein, Naeger, & Maass, 2005) using the UDF model. The arrival of a spike k after interspike interval Δ_{k-1} induces an increase in the post-synaptic charge of amplitude A_k:

$$A_k = w \cdot u_k \cdot R_k \tag{3}$$

$$u_k = U + u_{k-1}(1-U)e^{-\Delta_{k-1}/F} \tag{4}$$

$$R_k = 1 + (R_{k-1} - u_{k-1}R_{k-1} - 1)e^{-\Delta_{k-1}/D} \tag{5}$$

where w is the weight of the synapse (synaptic efficacy), u_k and R_k are hidden dynamic variables maintaining the facilitatory and depressionary tendencies of the short-term plasticity of the synapse, and U, D and F are the parameters modulating synaptic use, time constant of depression (in seconds), and of facilitation. Initially, $R_k = 1$ and $u_k = U$. Each spike contributes A_k to its PSR at the time it hits.

Recurrent Circuit (Liquid)

The auditory recurrent circuit ("liquid") is a $15 \times 5 \times 5$ column of current-based leaky integrate-and-fire (LIF) neurons (for a total of 375 neurons; 20% are randomly chosen to be inhibitory).

For the neurons in the liquid, $I_{bg} = 13.5$ mV uniformly and $\tau_m = 30$ ms. V_{rest} is 0 mV. When the membrane potential of a neuron exceeds V_{thresh} (15.0 mV), the membrane potential is reset to V_{reset} (13.5 mV) and the neuron enters a refractory period during which its dynamics are frozen. For excitatory neurons this is 3 ms (inhibitory 2 ms). I_{syn} is equal to the difference of the post-synaptic responses (PSR) of excitatory afferent synapses and the PSRs of inhibitory afferent synapses.

The probability that a synapse exists between neurons at 3-D points a and b is $C \cdot e^{(-D(a,b)/\lambda)^2}$, where λ is a global parameter controlling the density of connections ($= 2.0$), $D(\cdot)$ is the Euclidean distance function, and C is a parameter to modulate the probability of a synapse depending on properties of the connected neurons. In our case, $C = 0.3$ if a is an excitatory neuron and b is an excitatory neuron (EE), $C = 0.2$ for excitatory and inhibitory neurons (EI), $C = 0.4$ for inhibitory and excitatory neurons (IE), and $C = 0.1$ for two inhibitory neurons (II).

The parameters (U,D,F) were selected for each synapse depending on the type of neurons that were connected and were drawn from a Gaussian distribution with means (0.5, 1.1 s, 0.05 s) for EE, (0.05, 0.125 s, 0.120 s) for EI, (0.25, 0.7 s, 0.02 s) for IE, and (0.32, 0.144 s, 0.06 s) for II (standard deviation 50% of the respective means). Negative results were redrawn from a uniform distribution between 0.001 of the mean and double the mean. The weights w of the synapses were drawn from Gamma distributions with means 30.0 (EE), 60.0 (EI), 19.0 (IE), and 19.0 (II); SD 100% of mean, with negative results redrawn from a uniform distribution as described above. In addition, a synaptic delay of 1.5 ms was implemented for EE synapses, 1.0 ms otherwise.

Input Neurons

Raw audio streams (PCM 16 kHz) are converted into firing probabilities by a cochlear model (Slaney, 1998) which approximates the instantaneous firing activity of the auditory nerve at different points along the cochlea ("cochlear input", gray neurons in Fig. 1). These probabilities are linearly scaled and injected as current into a set of "spike generating" LIF neurons ("spike generation", blue neurons in Fig. 1). There are three differently parameterized classes of spike generating neurons (the three columns of neurons), for onset, offset, and passthrough. The onset spike generating neurons have a low (no firing without input) baseline firing rate ($I_{bg} = 13.5$ mV), and receive strong positive input from the cochlear model in the form of the spike probability for that channel $\times$ 20000 nA (thus increasing activity when input is present in that channel). The offset spike generating neurons have a higher (firing without input) baseline rate affected by Gaussian noise $I_{bg} = 13.5 + \Gamma(3.0, 0.05)$ nA, where Γ indicates a value drawn from a Gaussian distribution mean 4.0 and standard deviation of 0.05. The input from the corresponding cochlear channel is scaled by -20000 nA, thus suppressing activity when input is present in that channel. The passthrough (direct) spike generation neurons have the same parameters as the onset spike generation neurons. $V_{thresh} = 15.0$ mV for all these neurons.

The actual onset and offset detector neurons (green input neurons in Fig. 1) receive dynamic synapses from the surrounding three spike generation channels of their corresponding class of spike generating neurons. These synapses modulate the current injected into the post-synaptic neuron based on pre-synaptic firing activity. Large pre-synaptic activity will cause an initial facilitation, followed by a longer depression in the strength of injected post-synaptic current per action potential. Thus, they will inject strong current for the first few pre-synaptic spikes, followed by less current for a period thereafter. This, combined with the different baseline

firing rates of the spike generators, is what implements the onset/offset detectors. The passthrough neurons have quickly-recovering dynamic synapses and perform more like static synapses, but limit their firing rate to a slower rhythm.

V_{thresh} for each sensitivity level of onset/offset/passthrough detector is:

$$V_{thresh} = V_{reset} + E_0 \cdot (D^i \cdot (c+1)) \quad (6)$$

where $E_0 = 1.0$ for onset/offset detectors and $E_0 = 0.2$ for passthrough neurons. $D = 1.414$, with i from $c = 0$ to $c = N$ for each of the N sensitivities of onset/offset detectors ($N = 1$ for the experiments, i.e. only one sensitivity level for onset/offset detectors). For the one passthrough level, D is scaled by a factor of 9.0.

For the dynamic synapses between the spike generation and the onset/offset/passthrough neurons, the UDF parameters are $(0.5, 1.1, 0.05)$ and $w = 3.0$ (onset), $(0.5, 0.025, 0.5)$ and $w = 9.0$ (offset), and $(0.5, 0.025, 0.5)$ and $w = 9.0$ (passthrough).

Input (offset/onset/passthrough) neurons synapse into a randomly selected 30% of circuit neurons via static synapses. The weight A of each of these input synapses is drawn from a Gamma ($shape = 1$) distribution with mean $A_{mean} = 18.0$ when the post-synaptic neuron is excitatory and $A_{mean} = 9.0$ when it is inhibitory. Negative weights are set appropriately from a uniform distribution between $0.001 \cdot A_{mean}$ and $2 \cdot A_{mean}$.

The cochlear model described in (Slaney, 1998)[1] was modified and updated to run in real time. Parameter defaults are retained (except for: $breakf = 500$, $qconst = 8.0$, $stepfactor = 0.5$, $sharpness = 5.0$, $notchoffset = 1.5$, $preemphfreq = 300$, $taufactor = 3.0$) producing 52 output channels which encode on each simulation step the probability that a spike occurs in that channel on that time step.

Readout Neurons and Phrase Integration

A final set of neurons (readout neurons) serve as classifiers (r_n, one for each category n). They receive as input a weighted projection of the liquid's instantaneous firing activity (if spiked $+1$, otherwise -1), low-pass filtered to mimic the change in post-synaptic membrane potential had the readout been modeled as an LIF neuron (time constant 30 ms). A readout neuron is said to fire at a given time point if the sum of its inputs exceeds a threshold (the bias term determined by the linear regression below). The shape of the weighted projection is determined by supervised learning on a training corpus for each readout neuron independently. This is achieved by linear regression of the matrix of the liquid response to all stimuli (with an additional bias column which is always -1), with a supervisor vector which contains $+1$ for every time point during which input was from the target word class, and -1 otherwise.

Phrase Integration is performed by injecting 1.0 nA per r_n spike into a corresponding readout integration neuron i_n (with membrane time constant $\tau_m = 50$ ms). The readout integration neuron is considered to be active when $V_m > 0.25$, with no reset or refractory period. These readout integrators provides a more continuous picture of the readout activity that is robust to small recognition errors.

Utterance onset and offset detection (modeled as a neuron with membrane potential V_{utter}) is performed by combining input from the onset/offset/passthrough input neurons with the current readout firing activity. Each spike of an onset/offset/passthrough neuron imparts directly $1.0/-0.5/0.7$ mV to V_{utter}, which decays exponentially with $\tau = 200$ ms. For a word onset to be detected, $V_{utter} > 1.0$, and the highest value of all n, $i_n > 0.25$. An offset occurs when either of these variables falls below the threshold. On onset, an accumulator neuron a_l increases its voltage at a constant rate of 0.002 mV/ms, with a threshold of 0.3 mV. An utterance is only considered to be an instance of a category (i.e. not noise) if a_l is over threshold. The accumulator is reset to 0 mV at offset. Another set of accumulators a_n(one for each readout integrator i_n) sums the value of its corresponding readout integrator i_n from the point an onset is detected, and are reset at offset.

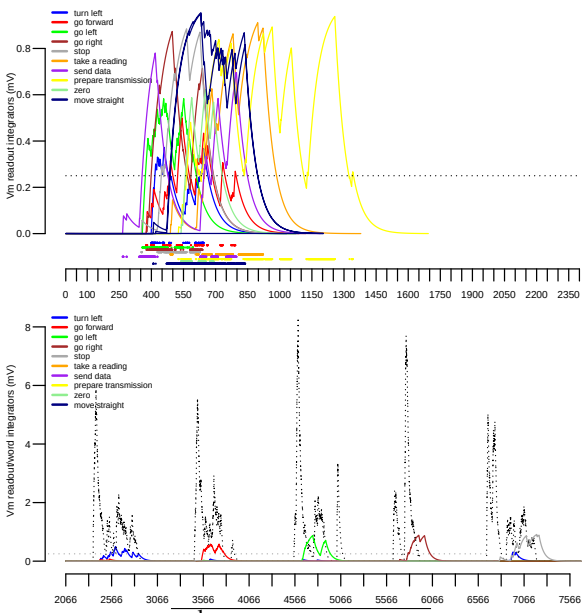
When an utterance that meets all the above conditions is detected, its category is determined by dividing each readout accumulator by the length accumulator. If the highest value is greater than the i_n threshold (0.25), the utterance is classified as being an instance of the winning category n. A token (e.g. a textual representation of it, or a number) indicating that category is returned.

Experimental Validation, Analysis and Results

The model was tested on part of the "CReST corpus" developed from human-human interactions in a search task (K. Eberhard, Nicholson, Kuebler, Gundersen, & Scheutz, 2010) (`http://www.cs.indiana.edu/~riveale/hricorpora.html`). The liquid was trained 10 times each on 9 audio samples each for each of ten phrase categories. In addition, it was "counter-trained" on a sample of recorded microphone noise (that portion of the supervisor vector for all categories was -1 for all time-points for the linear regression). The final two audio samples for each category were set aside for testing.

To test phrase recognition, a new audio stream was created by concatenating the 10 test samples which were presented to the model in real-time. The returned categories were verified to be recognized only once during the correct portion of the audio stream. The best liquid was able to correctly recognize all ten phrase categories. 50 liquids were randomly generated and trained. All were able to recognize at least 7 of the categories, with the exception of 3 liquids that only recognized 5 of them (recall that liquids are randomly generated). By observing the readout behavior during recognition, it was determined that the primary reason for failure was similarities in the categories (e.g. it was most likely to fail when phrases shared similar words, or words had large regions that were similar-sounding).

[1] Code from `http://www.slaney.org/malcolm/pubs.html`



class	acc_{on}	acc_{off}
0	0.18	1.00
1	0.27	1.00
2	0.41	1.00
3	0.55	1.00
4	0.60	1.00
5	0.60	0.98
6	0.55	0.99
7	0.73	1.00
8	0.33	1.00
9	0.72	1.00

Figure 2: *Top*: Readout (spike) and integrator (lines) response of each trained category from the best liquid to a test case of its own class. *Center*: Readout integrator responses of 5 phrases uttered sequentially, along with scaled utterance detector. Horizontal line is threshold. Note each "correct" readout is active during its own category and relatively inactive otherwise. *Bottom*: Proportion of correctly spiking and correctly not spiking time points for each trained readout (corrected for large leading/trailing silence).

Figure 2 shows the time-line of phrase recognition over the course of the test speech stream in two different ways (including the raw spiking characteristics of the best liquid for the auditory corpora used). Note that even *during* an utterance, it can be predicted based on readout integrator levels whether the utterance stream is probably of a given class. This information can be used to prepare (pre-cache or "prime") responses to that phrase.

The model has also recently been used for speech recognition embedded in a larger cognitive model of embodied situated human natural language interactions in human-robot interactions tasks, thus providing evidence for its utility in real environments (video at `http://www.cs.indiana.edu/~riveale/muridemo.html`). In this task, the model played the role of a speech recognizer that sends processed tokens representing whole phrases to a natural language parser and understanding system. This system processed the meaning of the phrase, and passed it on to a cognitive architecture (planner, etc.) which was able to initiate actions (such as changing its own goal state, change its behavior, learn new actions such as door-pushing) based on the content of the phrases (Cantrell, Talamadupula, Schermerhorn, Benton, & Scheutz, 2012).

Discussion

While previous research has presented methods for converting spike-encoded streams (generated offline from audio input) into liquid activity and, subsequently, liquid activity into instantaneous category readout firing, the problem of using readout firing activity to generate category-tokens of the kind used by most higher-level cognitive models has not been sufficiently addressed. Moreover, the method of encoding sound as spikes did not take into account the onset/offset-detection capabilities of the human auditory system, which contributes significantly to the robust recognition of shorter phonemes like consonants. The model presented in this paper addresses both shortcomings by processing sound in a realistic way up to the neural readout-level and being capable of returning correct word tokens based on readout activity at the right time. One limitation of the current model that must be addressed in the future is that some aspects of the token conversion (the conversion from raw spiking activity to phrase-tokens) are not guarenteed to work well in all speech situations. For example, the assumption that phrases will be preceded and followed by silence may not hold in situations where speech is very fast. In those situations recognizing phrases based on other information such as prosody could lead to better results (Christiansen, Allen, & Seidenberg, 1998).

In addition to the practical benefits of having a real-time model that shows promising performance on natural speech, the model makes important predictions about the mechanisms by which humans are capable of extracting discrete category information from a continuous real-time stream of sensory data. For example, the model proposed an explanation for how category priming or biasing effects come about, as the ongoing activity of each readout integrator can be viewed as the probability that the category it represents is currently present in the stimulus stream. This probability can be made to bias ongoing behavior even before the stimulus signal ends. To see this, imagine an experimental paradigm (e.g., the visual world paradigm (K. M. Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995)) in which an auditory cue (e.g., a color word) determines to which of two locations the subject should direct her eyes – it is well-known that humans in that case are capable of performing eye saccades shortly after the onset of the color word (even though they will not always perform them right away). For computational models of these human behaviors it means that a decision for eye movements has to be reached before the word stimulus ends. Specifically, if there are two category readouts, one for "Red" and one for "Blue", and the first phoneme of "Red" is presented, the model could use the already high activity of the "Red"

category to bias its looking behavior even before the stimulus has ended and the category word "Red" is fully recognized.

This example demonstrates that there can be multiple routes from the "readout" level to other parts of the cognitive system that can influence and bias behavior (one being the route where recognized word tokens are passed on to higher syntactic and semantic processing areas, while another being the more direct, intermediate route that can bias looking behavior). The current model is silent about the exact nature of these routes, as this would require additional specifications of those higher-level cognitive components and behavior-generating components, an important direction for future work. The current model is also silent about the important top-down biases coming from various other parts of the human cognitive system such as the syntactic and semantic biases as well as other biasing information based on perceptual, dialogue, task, and goal information. These top-down biases are critically involved in human speech processing and contribute to the robustness of human speech recognition in noisy environments. However, we believe that the particular model architecture will directly allow for this kind of information integration by way of appropriate top-down connections to the readout units whose activations represent the dynamically changing hypotheses about recognized words. We are currently investigating extended versions of the model that include additional perceptual areas (e.g., as described in (Veale, Schermerhorn, & Scheutz, 2011)) to test the extent to which perceptual biases can influence recognition rates and behavior (such as eye saccades).

Different from our previous models (Scheutz, Eberhard, & Andronache, 2004) which only at a high level of abstraction resembled the human cognitive architecture and only modeled the human data qualitatively, the goal for these new models is to be fully realized in neural architectures and to model the human data quantitatively.

Conclusions

In this paper, we presented a novel biologically plausible model for human speech recognition together with a proof-of-concept implementation and evaluation of the model. As part of the model, we introduced new methods for any-time phrase recognition based on the human early auditory system coupled with biologically plausible methods to produce word category tokens from a continuous auditory stream. We also introduced biologically plausible implementations of onset/offset detectors for word signals. Future work will include the already mentioned extensions by perceptual areas to be able to allow for biologically plausible quantitative neural models of human eye gaze behavior during reference resolution. In addition, we will also investigate mechanisms for improving noise-robustness with multiple sensitivity levels at the onset/offset stage and online learning of novel word categories.

Acknowledgments

RV is an NSF IGERT and NSF Graduate Research Fellow.

References

Cantrell, R., Talamadupula, K., Schermerhorn, P., Benton, J., & Scheutz, S. K. M. (2012, March). Tell me when and why to do it! Run-time planner model updates via natural language instruction. In *Proceedings of the 2012 human-robot interaction conference* (p. forthcoming).

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2/3), 221-268.

Eberhard, K., Nicholson, H., Kuebler, S., Gundersen, S., & Scheutz, M. (2010). The indiana cooperative remote search task (crest) corpus. In *Proceedings of lrec 2010: Language resources and evaluation conference.* Malta.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409-436.

Ghitza, O. (1987). Auditory nerve representation criteria for speech analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *35*(6), 736-740.

Legenstein, R., Naeger, C., & Maass, W. (2005). What can a neuron learn with spike-timing-dependent plasticity? *Neural Comput.*, *17*(11), 2337–2382.

Lyon, R. F. (1982, May). A computational model of filtering, detection, and compression in the cochlea. In *Acoustics, speech, and signal processing, ieee international conference on icassp '82* (p. 1282-1285).

Maass, W., Natschlager, T., & Markram, H. (2002). Real–time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, *14*(11), 2531-2560.

MacLeod, K. M., Horiuchi, T. K., & Carr, C. E. (2007). A role for short-term synaptic facilitation and depression in the processing of intensity information in the auditory brain stem. *Journal of Neurophysiology*, *97*, 2863-2874.

Scheutz, M., Eberhard, K., & Andronache, V. (2004). A real-time robotic model of human reference resolution using visual constraints. *Connection Science Journal*, *16*(3), 145–167.

Slaney, M. (1998). *Lyon's cochlear model* (Tech. Rep. No. 13). Apple Computer Inc. Cupertino, Ca.

Smith, L. S., & Faser, D. S. (2004). Robust sound onset detection using leaky integrate and fire neurons with depressing synapses. *IEEE Transactions on Neural Networks*, *15*(5), 1125-1134.

Veale, R., Schermerhorn, P., & Scheutz, M. (2011). Temporal, environmental, and social constraints of word-referent learning in young infants: A neurorobotic model of multimodal habituation. *IEEE Transactions on Autonomous Mental Development*, *3*(2), 129-145.

Verstraeten, D., Schrauwen, B., & Stroobandt, D. (2005). Isolated word recognition using a liquid state machine. In *Esann* (p. 435-440).