Linking Cognitive Tokens to Biological Signals: Dialogue Context Improves Neural Speech Recognizer Performance

Richard Veale (riveale@indiana.edu)
Indiana University, 841 Eigenmann Hall
Bloomington, IN 47406 USA

Gordon Briggs (gbriggs@cs.tufts.edu)
Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Matthias Scheutz (matthias.scheutz@tufts.edu)
Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Abstract

This paper presents a hybrid cognitive model engaged in experiments demonstrating a successful mechanism for applying top-down contextual bias to a neural speech recognition system to improve its performance. The hybrid model includes a model of social dialogue moves, which it uses to selectively bias word recognition probabilities at a low level in the neural speech recognition system. The model demonstrates how symbolic and neurologically inspired components can successfully exchange information and mutually influence their processing. Furthermore, the biasing mechanism is grounded in brain mechanisms of perceptual decision making.

Keywords: Speech Recognition; Liquid State Machine; Dialogue Context; Top-Down Bias; Signal-to-Token Conversion

Introduction

Human cognition comprises high-level knowledge-based processes as well as low-level perceptual and motor processes, both of which are implemented via electro-chemical mechanisms in the brain. High-level cognitive processes are often viewed as symbolic and discrete, while low-level perceptual and motor processes are subsymbolic and continuous. Moreover, high-level processes are taken to operate on structured representations, while low-level processes will usually not be representational at all. Two key challenges in cognitive science are thus to understand (1) how high-level processes are realized in "neural hardware" and (2) how they can interact with low-level processes (e.g., how discrete symbolic knowledge can influence continuous subsymbolic processes and vice versa). We will focus on the second challenge in this paper.

Connectionist computational modeling has made significant progress in addressing (1) over the years, producing more and more refined neurologically plausible models of cognitive functions which are verified physiologically (e.g. (Machens, Romo, & Brody, 2005)). However, fewer efforts have been made to address (2). Only recently, hierarchical Bayesian models have been proposed as a natural, systematic way to connect higher-level to lower-level processes (Kemp & Tenenbaum, 2008). Similar to the Bayesian approach, our goal is to understand the interactions between these two types of processes which operate at fundamentally different levels.

Hierarchical Bayesian modeling often focuses on the "computational level" (Marr, 1982), showing how higher-level processes can influence lower levels (e.g., by showing how distributions of higher-level structures constrain distributions of lower-level items). In contrast, our approach attempts to address all three levels and their mutual interactions. This is because these levels cannot be considered in complete isolation in cases where higher-level processes have to interact with lower-level processes in real-time contexts with real-world inputs. Specifically, we claim that the nature and time-course of low-level processes imposes significant constraints on the possible ways of exchanging information with higher-level processes. Low-level processes will limit the types of computations that are allowed in higher-level processes that communicate with them, since they may have stringent timing requirements and will not wait for a computation to finish with a result. Proposals that do not incorporate those constraints might result in models that produce correct results under some empirical regimes, but which are infeasible given implemenation constraints.

For example, a hierarchical Bayesian model of natural language processing might be able to show that high-level knowledge about grammar can successfully bias low-level speech processing, but whether that particular computational way of biasing is actually feasible and realistic in humans can only be determined by taking algorithmic and implementation constraints into account. These constraints include time bounds caused by the incremental nature of the speech processor. In this case the high-level computation can not expect to have access to a whole utterance before it starts biasing, since by that point the speech processor will already have advanced past the point where it is useful. Thus, although there are many ways in which higher levels could influence lower levels at the computational level, most of them are not realized in humans because of implementation or algorithmic constraints.

This paper makes three contributions: first, we will present a general way of integrating high-level processes operating on structured symbolic knowledge with low-level neural pro-

cesses with unstructured signals; second, we will show in the specific context of real-time biologically plausible speech recognition how high-level knowledge about diologues and mental states of interlocutors can be used to dynamically adjust parameters in the neural speech recognizer to improve recognition performance; and third, we will provide results from a real-time evaluation of the implemented model. The model includes a biologically plausible neural speech recognizer, a statistical/symbolic natural language understanding system, and a logic-based model of pragmatical and mental state inference. Previously, we have addressed the bottom-up transfer of information, i.e., conversion from the continuous stream of auditory neural firings to symbolic word tokens expected by a natural language processing system (Veale & Scheutz, 2012b). In this paper we address the reversed direction, the *top-down* transfer of information and biasing of low-level processes. Specifically, high-level knowledge-based representations of dialogue and interaction context will be used to bias auditory neural activity to improve word recognition performance in spoken language dialogues.

Background

In humans and other animals, perceptual decisions are modulated by system state in a top-down manner. Top-down biases have been documented empirically in a variety of contexts such as *vision search* (Chen & Zelinsky, 2006), *perceptual decision about motion* (Hanks, Mazurek, Kiani, Hopp, & Shadlen, 2011), *auditory disambiguation* (Hannemann, Obleser, & Eulitz, 2007)), and others. Furthermore, we are beginning to understand the mechanisms underlying these biases thanks to a combination of neurophysiological studies and behavioral research (e.g. see (Hanks et al., 2011). Perceptual decisions can be well-modelled using parallel diffusion processes (Ratcliff, Gomez, & McKoon, 2004), and there is evidence that these processes are realized in the brain as neural integrators collecting evidence for each alternate hypothesis independently. Prior probabilities influence the neural integrators based on the past experience of the organism. These influences have been shown to be caused by top-down biases, although some evidence exists that sensory cortex parameters also adapt to match environmental priors (Fiser, Chiu, & Weliky, 2004), which are outside the scope of this paper (Veale & Scheutz, 2012a). The shape and parameters of the thresholds and the bias functions responsible for top-down biases on behavior are still under active investigation (Hanks et al., 2011). However, the detailed behavior of these processes is not necessary to implement a working model that takes advantage of the general mechanism of top-down bias to improve perceptual decisions.

In this paper we are specifically interested in top-down biases on auditory word recognition. Contextual biases on word recognition are ubiquitous in the everyday world. For example, visual context and gesturing can be used in noisy situations to produce a sensible hypothesis for what a speaker is saying. This is not a novel observation. Top-down biasing of speech recognition probabilities have been investigated in a traditional speech recognition system (e.g. (Young, Hauptmann, Ward, Smith, & Werner, 1989)). Our work differs from this previous work in that the speech recognition system is built of biologically-plausible neural circuits modelling the early human auditory system. Although the general concept of using context to bias state in the speech recognizer is similar, the non-symbolic nature of the speech recognizer in our system requires serious reconsideration of how to actually implement the top-down bias. In this paper we adopt a simple approach and bias the temporal integrators representing the competing word categories, which directly influences the symbolic output of the speech recognizer.

The next section presents a short overview of the two most relevant portions of the hybrid model used in this paper. It describes the mechanism for top-down biasing of the neural speech recognizer, and overviews how the system operates.

Model Overview

The architecture of the cognitive model used for the experiments in the Experiment Setup Section is summarized in Figure 1. The neural speech recognizer (LSM ASR) is responsible for translating the acoustic signal into text tokens, which are sent to the NLP component. The NLP component parses the text tokens, and performs semantic analysis and utterance type classification. The dialogue system receives semantic information from NLP and updates the agent's beliefs, based on a pragmatic analysis (Briggs & Scheutz, 2011). The dialogue component also tracks the state of the current dialogue exchange, allowing for predictions about expected upcoming utterance types. Details of how biasing is implemented in the speech recognizer and Dialogue components are presented in the sections below. The model is implemented in the DIARC cognitive architecture (Schermerhorn et al., 2006), whose natural language capabilities have been demonstrated in human-robot interaction scenarios [1] (Cantrell, Scheutz, Schermerhorn, & Wu, 2010; Cantrell, Schermerhorn, & Scheutz, 2011; Briggs & Scheutz, 2012).

The Dialogue Component

The dialogue component contains knowledge of common dialogue exchange patterns, such as those in Table 1.

Table 1: Dialogue exchange patterns

Exchange Pattern	Dialogue Move Sequences
Statement-Ack Pair	$Stmt(\alpha,\beta) \rightarrow Ack(\beta,\alpha)$
Yes-No QA-Pair (pos)	$AskYN(\alpha,\beta) \rightarrow ReplyY(\beta,\alpha)$ $\rightarrow Ack(\beta,\alpha)$
Yes-No QA-Pair (neg)	$AskYN(\alpha,\beta) \rightarrow ReplyN(\beta,\alpha)$ $\rightarrow Ack(\beta,\alpha)$
QA-Pair	$AskWH(\alpha,\beta) \rightarrow Stmt(\beta,\alpha)$ $\rightarrow Ack(\alpha,\beta)$

[1] http://www.youtube.com/watch?v=RJ1VSIi1CM4

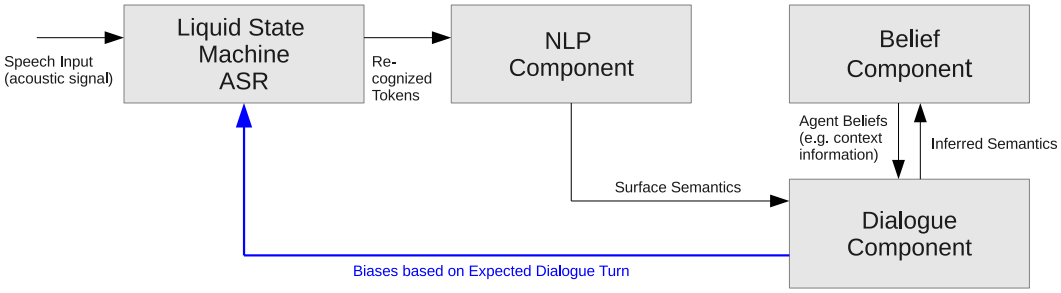


Figure 1: Information flow through the natural language system. The blue arrow indicates the top-down dialogue context bias on the ASR component introduced in this paper.

$Stmt(\alpha,\beta)$ denotes a statement utterance direct from agent α to agent β, while $Ack(\beta,\alpha)$ denotes an acknowledgment (e.g. "okay") from β to α. *AskYN* and *AskWH* denote a yes-no question and general question, respectively.

In this paper we focus on sending bias information to the LSM ASR component in the case of yes-no question-answer (QA) pairs. When the dialogue component detects a yes-no QA-pair has been entered, it sends a list of expected words to the LSM ASR component, specifically "yes" (*ReplyY*) and "no" (*ReplyN*). For each expected word x_i, a weight value $0 \leq w_i \leq 1$ is also sent to the LSM, denoting how much to weight x_i relative to other biased words (where 0 is equivalent to no bias and 1 indicates maximum bias).

The Speech Recognizer

The neural speech recognition system employed in this paper has previously been used to perform speech recognition for real-time human-robot interaction tasks (Veale & Scheutz, 2012b). The system converts from speech input streams to word tokens that can be used by other components of the cognitive model. The speech recognizer employs the liquid state machine (LSM) computational paradigm (Maass, Natschlager, & Markram, 2002) to perform recognition on audio input streams. The LSM is implemented using spiking neurons, and readouts are trained via linear regression. Figure 2 presents the main components of the speech recognizer.

Sound is processed into auditory nerve firings corresponding roughly to the strength of frequency channels in auditory input (Figure 2, left). These neurons project to several groups of pre-processing neurons (superior olivary complex) via groups of differently parameterized synapses, resulting in neurons sensitive to the onset/offset/passthrough activity for each cochlear channel. These pre-processing neurons in turn project randomly to the recurrent circuit (liquid), which is a large circuit of randomly connected spiking neurons. "Readouts" (discussed below) are trained via linear regression on a corpus of sound files, with supervisor vectors set to $+1$ for all instances of the target category and -1 otherwise. Additionally, all readouts are counter-trained against a "noise" corpus in which every readout's supervisor vector is -1.

Signal-to-Token Conversion Readouts (perceptrons) are trained via linear regression to respond positively to liquid activity patterns similar to liquid activity patterns evoked by the word examples they were trained on. Readouts are integrated over time with exponential decay (low-pass filtered, time constant 20 ms), and the value of these are continuously summed into the diffusors (right). In the model, readouts, integrators, and diffusors are only updated every 20 ms. The value of the readout integrator for readout r, σ_r is thus defined by the following equation (where τ_σ is the time constant and I_r is the input from the corresponding readout):

$$\frac{\partial \sigma_r}{\partial t} = \frac{-\sigma_r}{\tau_\sigma} + I_r \tag{1}$$

The diffusors compete with one another proportional to how strong their input is. The value of readout r's diffusor, Δ_r, is updated according to the following rule:

$$\Delta_r(t) = (\Delta_r(t-1) + \sigma_r) \cdot \frac{\sigma_r}{\sum_j (\sigma_j)} \tag{2}$$

This mechanism prevents the diffusion processes of ambiguous words from reaching threshold simultaneously. Using this system, there must be the equivalent of 100 ms of strong unambiguous evidence for a particular word category before it crosses threshold. This evidence could be provided by longer but weaker evidence, or by top-down bias.

Biasing Mechanism The biasing mechanism functions by injecting energy into the readout integrators, i.e., one level before the diffusion processes. The biaser specifies which categories should be biased, and the relative strengths for those biases. In the current paper, the amount of energy injected with a unit strength of 1.0 is equal to amount that is injected when the corresponding readout is active, thus up to "doubling" the input to the integrator at times when its presynaptic readout is active. Note that this implements the "simplest" diffusion model bias, involving linear bias to the diffusor's input diffusing to a constant threshold.

The result of bias is that biased words have "stronger" responses from their internal integrators, which translates to greater force of growth towards the diffusion threshold. This results in both faster reporting of the word (when the diffusor crosses threshold), and also stronger "confidence" in the word when the words offset is reported at the end of the word.

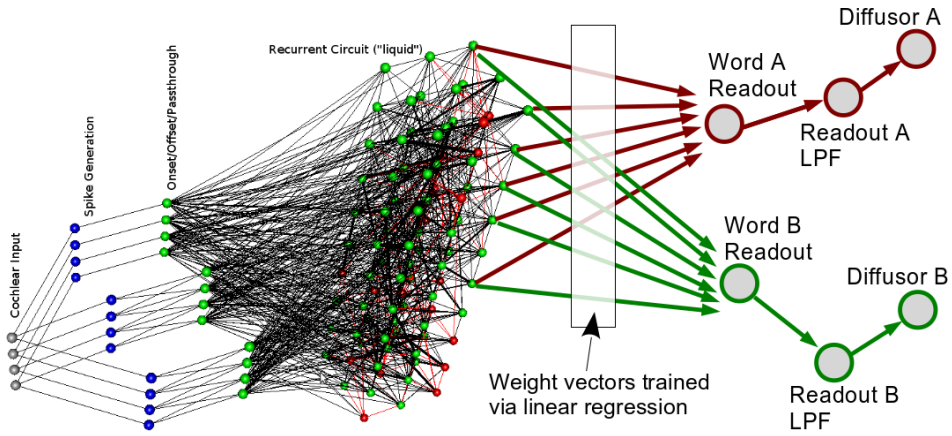


Figure 2: Visualization the neural model described in this paper. The pictured circuit has only 4 input channels, and a $3 \times 3 \times 10$ recurrent circuit. The actual circuit has 84 input channels and a $5 \times 4 \times 20$ recurrent circuit.

The LSM ASR was trained on five spoken instances of eight different words from the same speaker: *yes, no, guess, bess, jess, joe, bob* and a null response (background noise). The audio files used for testing are the same words spoken by a different speaker of the same gender. The words were chosen because several rhyme or have similar phonetic components to the "target" words "yes" ("guess", "bess", "jess") and "no" ("joe"), or share none ("bob").

The scenario we examine in this study consists of a simple yes-no QA-pair. The system is initiated with an intention to know whether its interlocutor possesses a particular mug in the belief component. The dialogue component, which queries the belief component for intentions to know information, generates the appropriate yes-or-no question:

```
Robot:  Do you have the mug?
```

After this NL reuest is generated, a response audio file is presented to the system. These audio files consist of "yes" and "no" responses recorded from a *different* speaker. Four conditions were examined: (1) "Yes" response, no bias; (2) "Yes" response, with bias; (3) "No" response, no bias; (4) "No" response, with bias. Data from the LSM (integrated readout activity and word recognition score) was recorded at 10 millisecond intervals over the duration of the input.

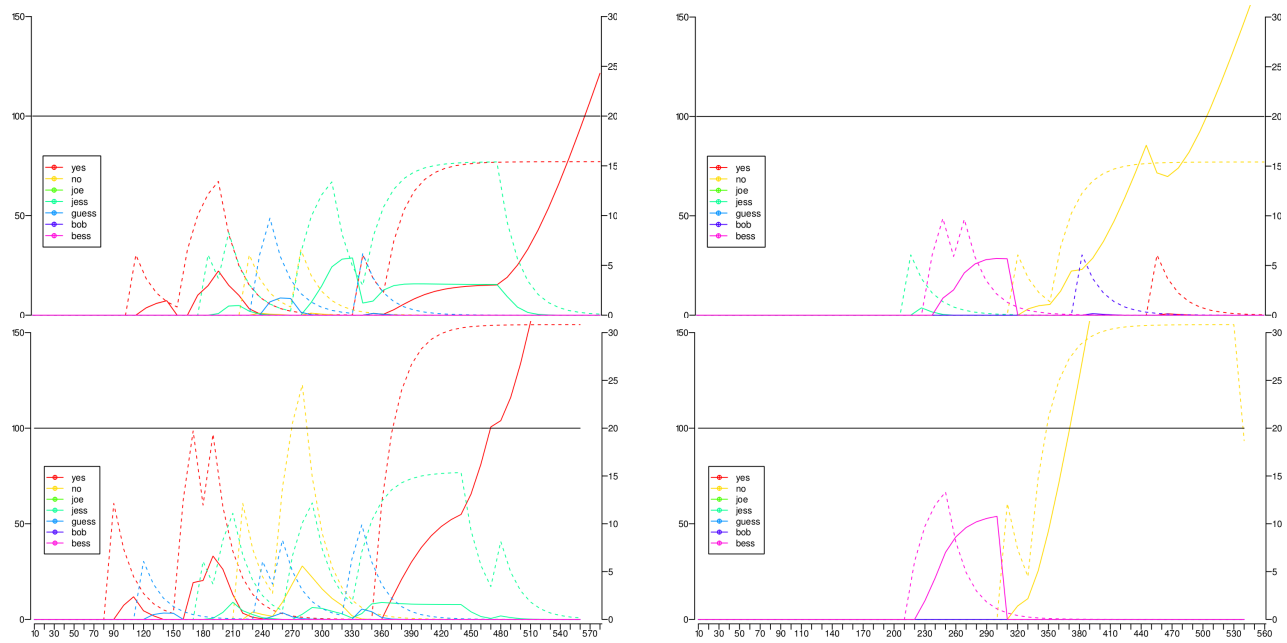
Results

The time course of the diffusors (solid lines) and readout integrators (dashed lines) for every word category are shown in Figures 3a (a "yes" trial) and 3b (a "no"trial). The primary comparison to make is the difference in the trajectories between the biased (each figure, bottom) and unbiased (each figure, top) trials. If the top-down biasing is working correctly, one should see a jump in activity over the unbiased trials for the contextually-appropriate words ("yes" and "no"), and no corresponding jump in any other words. This is precisely what is observed: even accidental weak responses to incorrect words ("bess" – purple in Figure 3b) do not seem to change significantly between biased and unbiased trials, whereas response to the appropriate word ("no", yellow) does. Similarly for Figure 3a, the activation of the contextually-inappropriate yet similar-sounding word "jess" (teal) does not change significantly between the biased and unbiased cases, yet the activation of the contextually-appropriate yet incorrect word ("no", yellow) is increased. Meanwhile, the activation of the contextually-appropriate and correct word ("yes", red) is stronger in the biased case and quickly advances to threshold.

As a control, a third set of experiments were run in which the responder responded with the similar-sounding but contextually-inappropriate word "joe" (Figure 4). In this case, the trajectories for all words do not differ significantly between the bias and unbiased conditions. However, in the biased condition (Figure 4, bottom), a slight jump in the recognition of the contextually-appropriate word "no"is seen near the end of the utterance. This is expected because the tail end of "no" is similar to "joe", and the additional contextual bias on "no" was sufficient to produce a small amount of drift in the diffusor for the period of similar sounds.

Experimental Setup

In terms of quantifying the advantage, one can look at the point at which recognition of the word reaches the confidence threshold (black horizontal bar). The diffusor in the "yes" unbiased condition (Figure 3a, top) crosses the recognition threshold at approximately 540 ms, whereas with bias the diffusor crosses the recognition threshold at approximately 470 ms (bottom), demonstrating a reduced recognition time. Note that the readout values for both "yes" and "no" responses are significantly increased in the biased condition compared to the unbiased condition, as both are anticipated as possible answers (whereas the readouts for the other word nodes remain relatively unchanged in amplitude). In the "no" unbiased condition, the diffusor crosses the recognition threshold at approximately 480 ms (Figure 3b, top), whereas with bias the diffusor crosses the recognition threshold at approximately 360 ms, again demonstrating a a reduced recognition



(a) Response to "yes" stimulus.

(b) Response to "no" stimulus.

Figure 3: LSM ASR responses to "yes" and "no" stimuli in no bias condition (top) and bias condition (bottom). The trajectory of activity for the readouts and diffusors for all trained words in response to the injected sound is plotted over time after the question is asked. Dotted lines represent the individual readout integrators for each word, while solid lines represent the diffusors. In both cases, the diffusor for the correct word (red solid line on left, yellow solid line on right) crosses the threshold significantly quicker in the bias condition (bottom). The influence of the top-down bias mechanism can be clearly seen in the increased activity of the readout integrators for "yes" and "no" (red and yellow dotted lines, respectively) in the bias condition.

time. Keep in mind that these are different words that begin at slightly different times and which extend for different amounts of time and have different volumes and distances from the training corpus. Thus it is important to focus on the differences *within* a word to see the performance increases resulting from top-down biasing.

Future Work

Expanding and refining the contexts in which top-down biasing of the speech recognizer will occur will provide ample opportunities for future research. For instance, incremental parse hypotheses in the NLP component could be used to identity likely upcoming words. Certain sentential modifiers (e.g. "I am *now* at the store" vs. "I am *still* at the store") can be used in conjunction with belief models and contextual knowledge for prediction purposes. If, for example, common ground in the dialogue exchange was established such that both speaker and listener knew the speaker was at the store previously. The partial sentence, "I am still at–" would be highly indicative of "..at the store". These semantic and belief model implications of these modifiers can be reasoned about in our pragmatics system (Briggs & Scheutz, 2011). Additionally, some yes-no questions are actually conventionally indirect forms of general questions. For instance, "Do you know who has the mug?" is often an indirect form of, "Who has the mug?" and may elicit a name in response. Our natural language system has mechanisms of recognizing and reasoning about such indirect speech acts (Briggs & Scheutz, 2013, forthcoming), and therefore more precise biasing algorithms are ripe for investigation.

A more theoretically interesting extension of the current work will more directly address the theoretical issues from the introduction. In the current paper, only pseudo-symbolic readout neurons were influenced by the top-down bias. This allowed us to explore the time-constraint theme, but not the disconnects in representation between multiple levels. In the future it will be interesting to directly bias the state of the auditory circuit, to further explore how such interactions could take place.

Conclusion

This paper introduced a hybrid neural-symbolic model that demonstrates not only the bottom-up communication of cognitive tokens from continuous sensory streams, but also the top-down biasing of neural speech recognition using predictions based on expected dialogue moves. The top-down biasing of the neural speech recognizer results in *faster* and *more confident* word recognition for contextually appropriate word categories during dialogue exchanges. The top-down biasing mechanisms are biologically accurate in that the effect of the top-down signal on high-level neurons in the speech recognition circuit parallels that observed in "diffusion" neurons

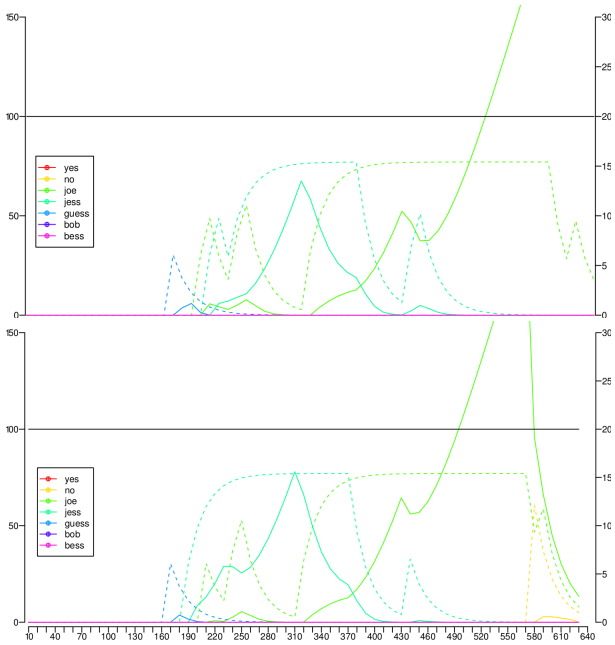


Figure 4: LSM readout results for "joe" response for no bias condition (top) and bias condition (bottom). Conventions are equivalent to figure 3. The diffusor for the actual uttered word "joe" does not significantly differ between the biased and unbiased conditions, crossing the threshold (black horizontal line) at roughly the same point in both conditions.

recorded from primate association cortex. The hybrid model presented in this paper engages interesting questions regarding interaction between different levels of abstraction. We use this to highlight that implementation-level details can actually constrain the computational level in real-time real-world situations. We believe that it is important to keep this relationship in mind when making claims about human cognition.

Acknowlegements

This research was in part supported by ONR grant #N00014-11-1-0493. RV is an NSF GRF and NSF IGERT.

References

Briggs, G., & Scheutz, M. (2011, June). Facilitating mental modeling in collaborative human-robot interaction through adverbial cues. In *Proceedings of the sigdial 2011 conference* (pp. 239–247). Portland, Oregon.

Briggs, G., & Scheutz, M. (2012). Multi-modal belief updates in multi-robot human-robot dialogue interaction. In *Proceedings of 2012 symposium on linguistic and cognitive approaches to dialogue agents.*

Briggs, G., & Scheutz, M. (2013, forthcoming). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the 27th aaai conference on artificial intelligence.*

Cantrell, R., Schermerhorn, P., & Scheutz, M. (2011, July). Learning actions from human-robot dialogues. In *Proceedings of the 2011 ieee symposium on robot and human interactive communication.*

Cantrell, R., Scheutz, M., Schermerhorn, P., & Wu, X. (2010, March). Robust spoken instruction understanding for HRI. In *Proceedings of the 2010 human-robot interaction conference.*

Chen, X., & Zelinsky, G. (2006). Real-world visual search is dominated by top-down guidance. *Vision research*, *46*(24), 4118–4133.

Fiser, J., Chiu, C., & Weliky, M. (2004). Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature*, *431*(7008), 573–578.

Hanks, T., Mazurek, M., Kiani, R., Hopp, E., & Shadlen, M. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience*, *31*(17), 6339–6352.

Hannemann, R., Obleser, J., & Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain research*, *1153*, 134–143.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *PNAS*, *105*(31), 10687–10692.

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.

Machens, C., Romo, R., & Brody, C. (2005). Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science (New York)*, *307*(5), 1121-1124.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt and Co., Inc.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159.

Schermerhorn, P., Kramer, J., Brick, T., Anderson, D., Dingler, A., & Scheutz, M. (2006). Diarc: A testbed for natural human-robot interactions. In *Proceedings of aaai 2006 robot workshop* (pp. 1972–1973).

Veale, R., & Scheutz, M. (2012a, November). Auditory habituation via spike-timing dependent. In *Proceedings of the international conference on development and learning and epigenetic robotics.* San Diego, CA.

Veale, R., & Scheutz, M. (2012b). Neural circuits for anytime phrase recognition. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (p. 1072-1077). Austin, TX: Cognitive Science Society.

Young, S., Hauptmann, A., Ward, W., Smith, E., & Werner, P. (1989). High level knowledge sources in usable speech recognition systems. *Communications of the ACM*, *32*(2), 183–194.