Integrating Cognitive Capabilities in a Real-Time Task

Greg Nelson
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213-3891
ghn+@cs.cmu.edu

Jill Fain Lehman
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213-3891
jef+@cs.cmu.edu

Bonnie E. John
Human-Computer Interaction Institute and
Computer Science and Psychology Departments
Carnegie Mellon University
Pittsburgh, PA 15213-3891
bej+@cs.cmu.edu

Abstract

NTD-Soar is a model of the perceptual, cognitive, and motor actions performed by the NASA Test Director as he utilizes the materials in his surroundings and communicates with others to prepare for a Space Shuttle Launch. The model, built within the framework of a serial symbolic architecture, is based on a number of independently designed general cognitive capabilities as well as a cognitive analysis of a particular task. This paper presents a detailed description of the model and an assessment of its performance when compared to human data. NTD-Soar's ability to display human-like real-time performance demonstrates that symbolic models with a serial bottleneck can account for complex behaviors which appear to happen in parallel, simply by opportunistically interleaving small elements of the different subtasks.

Introduction

Our goal in this research is to create an integrated cognitive model by combining several independently developed in-depth models of particular capabilities into a single agent. Specifically, we are modeling the behavior of the NASA Test Director (NTD). The NTD is responsible for coordinating many facets of the testing and preparation of the Space Shuttle before it is launched. He must complete a checklist of launch procedures that, in its current form, consists of 3000 pages of looseleaf manuals (the Operations and Maintenance Instructions, or OMI), as well as graphical timetables describing the critical timing of particular launch events. To accomplish this, the NTD talks extensively with other members of the launch team over a two-way radio net called the Operational Intercommunications System (OIS). In addition to maintaining a good understanding of the status of the entire launch, the NTD is responsible for coordinating troubleshooting attempts by managing the communication between members of the launch team who have the necessary expertise.

The complete model is made up of many capabilities which interact with a simulated physical world, including a simulated OMI and simulated OIS communications. Figure 1 shows an example of an OMI page along with a graphical representation of the model's eye movements during visual scanning. Our language model derives from NL-Soar's systems for comprehension (Lewis, 1993; Lehman et al., 1991) and generation (Rubinoff and Lehman, 1994). Decision making and problem solving knowledge in NTD-Soar came from an earlier NTD agent that provided a functional account of behavior in the domain (John et al., 1991). The visual processing capability was drawn from a number of sources, including the work of NOVA (Wiesmeyer, 1992). Each of these earlier models was built within the common Soar framework, which made it possible to import a great deal of this prior work directly. The general Soar architecture, which is independent of these models, is a classical, symbolic framework that allows limited parallelism in perceptual and motor processing but imposes a serial decision-making bottleneck in cognition (Laird et al., 1987). While serial models have been successful in areas such as planning and problem solving, there has been speculation that their serial nature makes them inadequate for real-time performance in other domains, including vision and language. In particular, Fodor (1983) claims that such systems will not be capable of handling real-time speech input or of simultaneously processing multiple senses. In counter-evidence to such claims, we present an implemented serial symbolic model that succeeds at these traditionally parallel tasks without violating the constraints of real-time processing.

The Origins of the Model

The NTD model is shaped by many forces: data gathered through psychological experiments, knowledge acquisition sessions with NTDs, NASA's standard audio recordings of launch communications, the physical artifacts in the NTD's environment, earlier models of individual capabilities, and the Soar architecture itself.

Much of the planning, problem solving, and other cognitive behavior was taken from a preliminary, purely functional model of the NTD (John et al., 1991). The cognitive processes in the preliminary NTD model were based on interviews, field observation, and analysis of the task, the artifacts, and the discourse. As a functional model, it made no attempt to account for timing data. In addition, this early model did not include the details of natural language comprehension, generation, or visual search.

The language comprehension capability of the current model, NL-Soar, was developed as part of separate, ongoing research on natural language in Soar (Lewis, 1993; Lehman et al., 1991; Rubinoff and Lehman, 1994). The development of the comprehension capability was guided in part by a wide range of psycholinguistic phenomena such as embedded constructions, non-problematic ambiguities, and garden-path sentences, directly comparing empirical results on the time course and success rate of comprehension for these different types of utterances with the NL model's predictions. A compatible model of language generation is being built within the established NL-Soar framework, and is used in NTD-Soar as well.

The organization of perceptual, cognitive, and motor actions of the NTD-Soar model takes it inspiration from the Model Human Processor (Card et al., 1983), that is, perception and motor actions proceed in parallel with serial cognition. Models describing parallel activities using this organization have been highly successful at predicting human behavior in a variety of Human Computer Interaction tasks (John, 1988; Gray et al., 1993; Chuah et al., 1994). Timing estimates for motor behaviors were based on human performance data on individual movements, including page turns (Egan, D. E., personal communication, March 19, 1991) and saccades (Westheimer, 1954; Fuchs, 1976; Card et al., 1983). The perceptual and cognitive aspects of visual attention were taken in part from the NOVA model (Wiesmeyer, 1992), which itself predicts the response times in a large set of classic psychological experiments.

A final contribution to the model comes from the Soar architecture itself, through its incorporation of many theoretically motivated assumptions about cognition. One assumption that is of particular importance for modeling the time course of behavior is the duration of a Soar *cognitive operator*, the smallest unit of serial cognitive behavior in Soar. Soar operators are considered within the Soar theory to correspond to a fixed duration of roughly 50 milliseconds of real time (John, 1988; Wiesmeyer, 1992; Lewis, 1993). This assumption provides the basis for all of the predictions of cognitive durations made by the system. Newell (1990) discusses the origin of this number, and its potential for variation. We believe that the best way to evaluate this assumption is for numerous researchers to utilize the same value in widely disparate domains, and have therefore chosen what we believe to be the most commonly used figure.

The focus of this research has been the integration of cognitive capabilities independently constructed in Soar. Relevant work outside of the Soar framework includes searching for information in structured texts (Carpenter and Alterman, 1994; Lohse, 1993), and real-time language processing for specialized domains (Torrance, in press; Horswill, in press; Traum et al., in press).

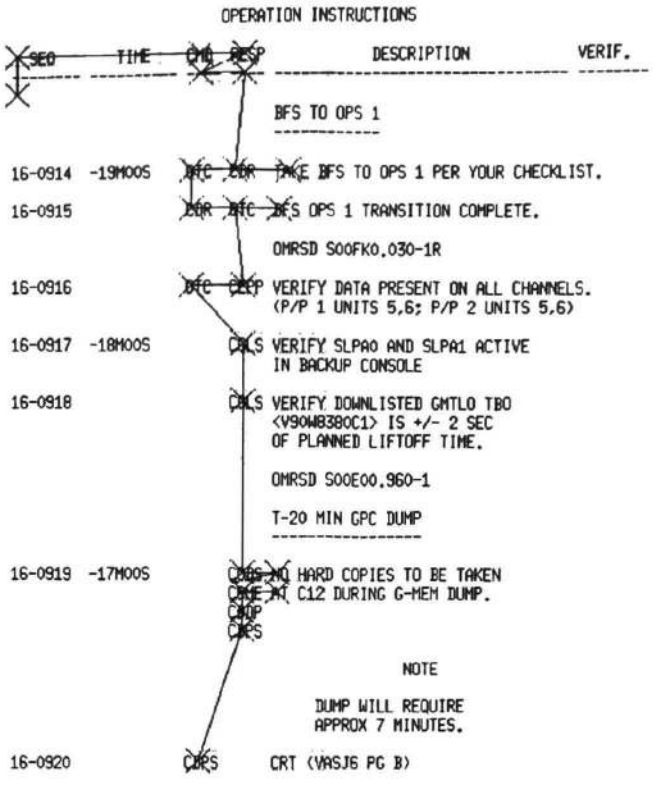

Figure 1: A sample page of the OMI showing modeling of eye fixations (boxes) and attention shifts (cross-hairs), during the NTD's search for his next participation. The focus of attention is on the two columns which provide information about the step participants.

The Integrated Model

In a Soar model, different types of knowledge are organized into *problem spaces*. Each problem space contains a set of *operators* that modify or change *states*. The operators are the smallest units of deliberate action and must be performed sequentially. States collect information that may come from the physical world through perception or from cognition, and may be thought of as the system's working memory. At any point, the knowledge required to unequivocally select or perform an operator may be unavailable in the current problem space; in this case, an *impasse* arises, and a subgoal is created to search for the knowledge in other problem spaces. When knowledge that is able to resolve the impasse has been found, the subgoal is accomplished, and a generalization of the problem solving in the subgoal is learned. This *chunk* integrates knowledge from one or more problem spaces and makes it directly available in the problem space which gave rise to the subgoal; thus in similar future situations, the system behaves *recognitionally*, i.e., without impasse.

NTD-Soar has many different problem spaces that reflect the origins of its knowledge. Language comprehension and generation problem spaces came from NL-Soar; task performance and selection spaces came from the earlier NTD work; and visual scanning spaces derive, in part, from NOVA. Apart from these problem spaces, a modest amount of additional *control* knowledge is needed to complete the model. This additional knowledge is necessary to mediate among capabilities contending for control of the central cognitive resource. This contention among capabilities did not arise in any of the prior models, because each one modeled only one capability. In order to achieve this integration, the knowledge in some of the prior systems, particularly in the earlier NTD work, had to be re-organized in a manner capable of supporting this sort of mediation and resource sharing.

Despite the relative separability of knowledge implied by the problem-space organization, the NTD's behavior consists of perceptual, motor, linguistic, and other cognitive tasks that are thoroughly intertwined. Often it is necessary that the tasks be performed in parallel, or at the very least give this appearance. Fodor (1983) argues that such behavior can be modeled only by the action of interconnected cognitive modules working in parallel. Others have focused on achieving parallelism by finding fixed ways to combine multiple parallel actions into a single larger action (e.g. Covrigaru, 1992). Our approach, however, is to explain the apparent parallelism as an ability to switch fluidly and opportunistically among tasks. Thus,

NTD-Soar achieves the appearance of doing several things in parallel by *interleaving* tasks, performing small elements of one task *between* elements of other tasks. Interleaving successfully substitutes for parallelism because the system as a whole is able to keep up with the various demands of the external environment.

The real NTD is an expert at his task. For this reason, much of the language and task behavior in the model should occur in Soar's *Top problem space*, which interacts with the outside world and where rapid, expert behaviors occur via knowledge that is available recognitionally. Recall that recognitional knowledge consists of associations in long-term memory that are immediately available because of the current contents of working memory, rather than those that must be found and integrated through deliberate search in subgoals. Soar's learning mechanism automatically builds this recognitional knowledge in its chunks, allowing novice models to develop into expert models. Since we wish to model expert behavior, our evaluations examine behavior after chunking has occurred.

In NTD-Soar, we have found it useful to categorize the Top-space operators in much the same way Newell did in his discussion of immediate behavior in Soar (Newell, 1990). These categories reflect the types of basic cognitive activities the NTD performs:

Attention operators direct the perceptual mechanism to provide additional information about particular elements in the sensory input.

Comprehension operators interpret and elaborate the perceptual inputs in conceptual terms, or perform further inferences on the conceptual representation.

Task operators select the next task to be performed, thereby biasing the selection of other operators.

Intend operators initiate physical responses from the system, such as simulated speech or hand movements.

All tasks are performed by collections of these types of operators, plus the perceptual and motor operators that interact with the physical world. Naturally, language comprehension is performed by perceptual and comprehension operators, and language generation leads to intention and motor operators. As Lewis discusses (1993), comprehension may take two to five cognitive operators per word in handling syntactic processing, semantic processing, and reference resolution. Generation involves a similar number of operators (roughly three) with discourse-level processing replacing reference resolution.

Processing of the visual page is also performed through similar operators – intention operators produce eye movements, and perceptual, attention, and comprehension operators search for information and perform the visual pattern matching that allows NTD-Soar to determine whether it has found the information it is seeking. The model predicts a series of eye movements, and within each eye fixation one or more shifts of attention, giving patterns such as the one shown in Figure 1. The amount of work that may be done by one attention operator is fairly well constrained by psychological data, as described by Wiesmeyer (1992). Visual scanning differs from reading because the "comprehension" process often does not need to consider either syntax or meaning; it is frequently possible to find the desired information simply by matching the pattern in visual attention to a target pattern.

The serial nature of the cognitive behavior and the time associated with the operators makes it possible to perform very direct comparisons between the model and human data. Further, the short duration of the operators (50 ms each) allows the predictions we make to have a very fine resolution, in many cases finer than the data we are modeling. In the remainder of this paper we explore the fit of the model to data we have, and discuss further predictions for which we have not yet collected the data needed to evaluate the model.

Performance of the Model

Communication in the launch room is done in a stylized, restricted form of English, somewhat akin to Seaspeak, the international language used aboard ships and aircraft, familiar in such terms as "Roger," "Copy," and "Niner" (Crystal, 1987). The performance data we have available from NASA contain only these communications, including the words spoken by the NTD himself. We have also collected eye movement data from well-practiced non-NTD participants scanning the OMI for launch information.

Segments of the NASA dialogs have been transcribed with detailed timing information, allowing us to evaluate how well we model the response times of an NTD to questions and other linguistic input. As a side effect, we can also look at the quality of our model with respect to other durations, such as the time it takes the NTD to find and perform his next step on the OMI checklist, by looking at the lengths of pauses between dialogs.

Table 1: Sample NTD Dialogs

Time (ms)	Speaker	Utterance
147357	CVFS	NTD, CVFS.
149686	NTD	Go ahead, CVFS.
150814	CVFS	Ready for BFS uplink.
152102	NTD	I copy.
152620	NTD	Houston Flight, NTD.
153969	NTD	Perform BFS preflight uplink loading.
156063	FLT	In work.

One of the dialogs we have explored is a communication between the NTD and the CVFS (another launch team member), which proceeds as shown in Table 1. The times associated with each line are measured in milliseconds from the start of the transcript two minutes earlier. The first utterance, "NTD, CVFS," is a request to initiate a dialog with a specific party (named first), made by the speaker (named second). The NTD replies with, "Go ahead, CVFS," to acknowledge that he is listening. The CVFS then provides the content of his communication: that the computer is ready to receive a radio transmission (uplink) of backup flight software (BFS). The NTD acknowledges this with, "I copy." The table continues with the next dialog, in which the NTD summons Houston Flight (FLT) to direct him to perform the uplink.

In order to model the timecourse of behavior, we simulated the passage of time, using the serial nature of the operators and

their estimated duration to provide a measure of accumulated time. We use this "clock" initially to determine the times at which the model hears the other speakers, and thereafter to measure when the system responds. We measured the duration of specific utterances from the audio tapes and used these as auditory input to the model. The same measured durations served as estimates of the duration of speech outputs from the model.

The audio data from the dialog in Table 1 and NTD-Soar's behavior are depicted graphically in Figure 2. The dotted boxes in the figure indicate the times of the observed speech in the human data, with the dotted lines connecting them to the model's perceptual input (black boxes) and its motor output (white boxes) for the corresponding events. Cognitive activities are shown in grey, and the discontinuities where one aspect of cognition (comprehension, generation, or page scanning) pauses briefly while another proceeds give a graphical depiction of the interleaving that is occurring. Each grey rectangle within the longer boxes represents a distinct cognitive operator, falling into one of the four categories of attention, comprehension, intention, or tasking. Preliminary analysis of eye movement data we have collected suggests that the model predicts the number and spatial distribution of eye fixations well, but it is impossible to display a one-to-one match between the model and real NTD behavior.

While no two cognitive events overlap, and motor activities using the same physical system (e.g. eyes, mouth) must be serial, there is, nevertheless, some parallelism exploited among cognition, perception, and motor activity. It is possible for the model to speak and move its eyes simultaneously, but these actions must be initiated sequentially by cognition in this model. A great deal of this interleaving goes on during the first dialog, while the NTD is trying to locate the next step in which he participates. Once he initiates the second dialog, the visual search task is no longer necessary, so he spends his cognitive time exclusively on communication.

As shown in Figure 2, the timecourse of behavior simulated by the model's speech output follows the performance data available from NASA quite nicely. This figure further illustrates how NTD-Soar manages its resources to keep up with external events over which it has no control. It receives auditory input from the external world at times and with durations specified by the observed data. To accomplish all the necessary tasks in the available time before more input begins, NTD-Soar interleaves scanning the page to find the next relevant step, comprehending the auditory input, and generating an appropriate response. The total time taken by the model from the when the CVFS hails the NTD until NTD-Soar informs Houston that they can perform the BFS uplink loading is 8392 msec, only 11% longer than the observed 7574 msec.

Although this prediction is quite good for a model of this type, accurately predicting the timecourse of behavior of *this particular NTD* during *this particular launch* is not the point of our modeling effort. The point here is to demonstrate a human-like model that keeps pace with the real-time constraints of a complex task despite being built within a symbolic architecture with a serial bottleneck in cognition. By "human-like behavior" we mean that both the content and timecourse of behavior displayed by the model are within the expected range for NTDs. In the stylized communications of the launch procedure, the content of the NTD's speech is often a identical to the words in the OMI, so the content of NTD-Soar's utterances is reasonable when it utters words from the OMI as we have implemented it to do. As for the timecourse, the pauses observed with the real NTD (after being hailed by the CVFS and after hearing "Ready for BFS uplink") are about 200 msec in length. In contrast, NTD-Soar pauses for 450 msec at the same places. However, empirical data show a range of mean pauses from about 400 msec for telephone conversations (Norwine and Murphy, 1938) to 700 msec for face-to-face conversations (Jaffe and Feldstein, 1970). Thus the pauses that NTD-Soar displays are well within the range of acceptability for human conversation.

It is important to note that this particular implementation of NTD-Soar is conservative in several ways, that is, it produces slower behavior than other reasonable implementations would produce. For instance, this implementation has a preference for listening to the entire summons ("NTD, CVFS") before generating any of the structure required for articulating the acknowledgement ("Go ahead CVFS"). A reasonable alternative strategy would be for NTD-Soar to generate the structure for "Go ahead" immediately after hearing "NTD," and generate the structure for "CVFS" after hearing those call letters (i.e., interleaving the comprehension and generation tasks) and then sequentially articulating the generated surface forms to output the entire sentence. This strategy would produce a shorter pause between summons and acknowledgement, closer to the behavior of this particular real NTD. Another strategy choice in this implementation that produces longer response times is that this NTD-Soar waits until an entire word (or acronym) is input before starting to comprehend that word. Thus, it waits until all four syllables of "CVFS" have been heard before beginning to comprehend who is speaking. This leads to a conservative estimate in that the speaker is uniquely identified by the first two or three call letters and may even be identifiable from voice alone with the first call letter. Furthermore, we are currently using a preliminary model of speech generation that we believe will require fewer operators per word as it is refined. For all these reasons, we find the current demonstration of human-like real-time behavior extremely encouraging, since the response rate will only get better with refinement of this model.

Currently, we can evaluate our model only by matching it to the data available from NASA, the audio tape of a launch, or by comparing it to our eye movement data. However, this model makes other testable predictions for which we hope to collect data in the future. Other observable actions not discussed here are also incorporated in the full model, e.g., turning pages and checking off steps. Collecting these data and comparing them to the predictions of the model remains for future work.

Conclusions

We have described a serial, symbolic model of the NASA Test Director that displays human-like, real-time performance by interleaving tasks in reaction to a simulated physical world. It makes testable predictions about the timecourse of behavior and fits reasonably well to the available verbal behavior observed in a real Space Shuttle launch. We believe that NTD-Soar demonstrates the plausibility of a symbolic architecture

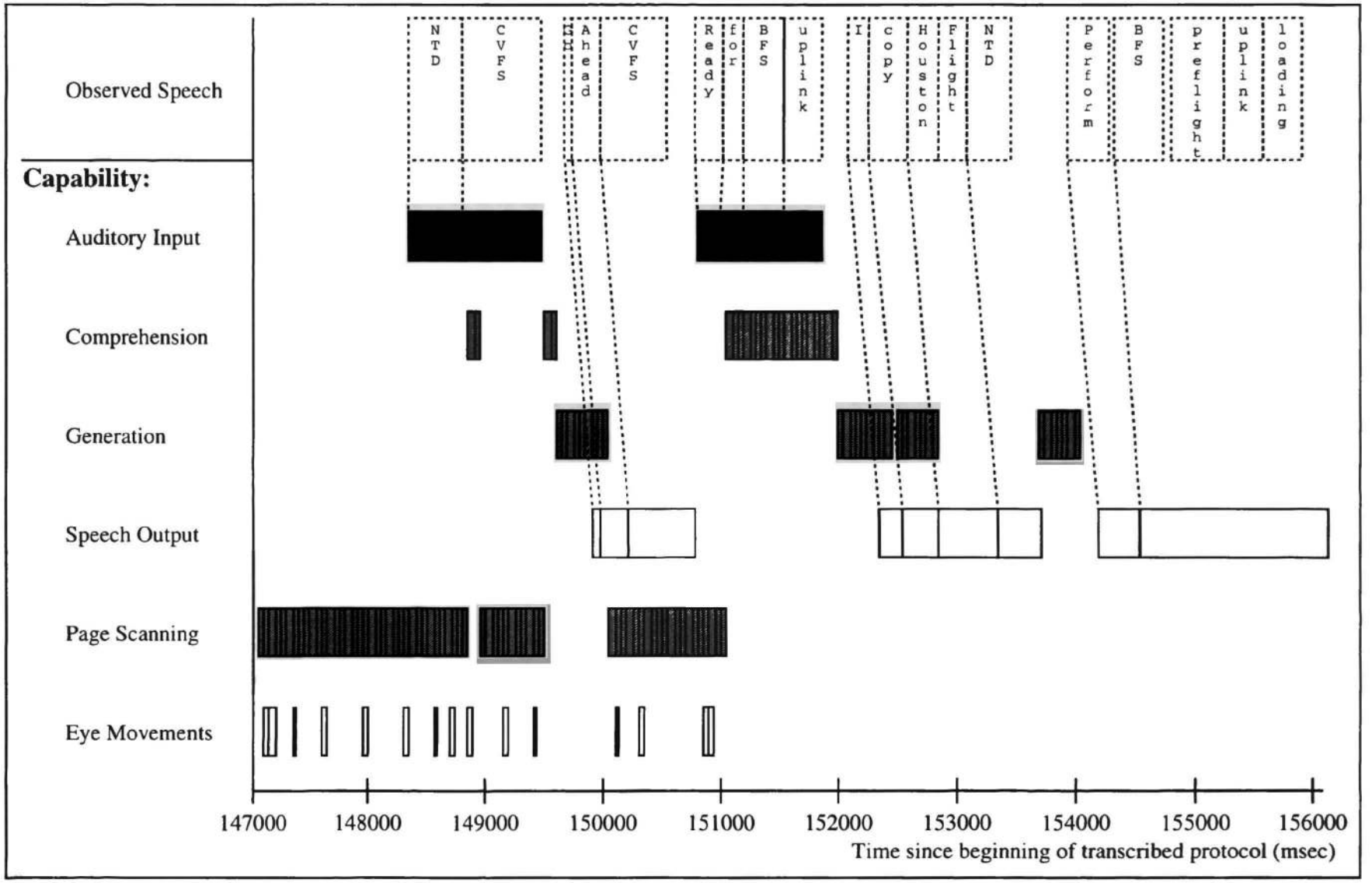

Figure 2: Graphical comparison of audio data and NTD-Soar's behavior.

Dotted lines indicate match with data, with slant reflecting discrepancies. Black boxes are perception of speech input; grey are cognitive operators; white are motor operators that affect the physical world.

with a serial bottleneck in cognition for modeling human behavior in a rapidly changing world. At the very least, such architectures cannot be dismissed by the claim that they cannot keep up with external events without some parallelism in cognition.

Acknowledgements

This research was supported in part by NASA, Grant No. NAG 2-634, in part by the Office of Naval Research, Cognitive Science Program, Contract Number N00014-89-J-1975N158, and in part by the Wright Laboratory, Aeronautical Systems Center, Air Force Materiel Command, USAF, and the Advanced Research Projects Agency under grant number F33615-93-1-1330. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of NASA, ONR, Wright Laboratory, or the U.S. government. This manuscript is submitted for publication with the understanding that the U. S. Government is authorized to reproduce and distribute reprints for Governmental purposes. We wish to thank Richard Lewis for extensive help and feedback during the integration of NL-Soar, Robert Rubinoff, who is working on generation within NL-Soar, and Roger Remington, who provided many valuable insights into visual attention.

References

Card, S. K., Moran, T. P., and Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carpenter, T. and Alterman, R. (1994). A taxonomy for planned reading. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Chuah, M. C., John, B. E., and Pane, J. (1994). Analyzing graphic and textual layouts with goms: Results of a preliminary analysis. In *Proceedings Companion of CHI, 1994*, pages 323–324. New York: ACM.

Covrigaru, A. (1992). *Emergence of Meta-Level Control in Multi-Tasking Autonomous Agents*. PhD thesis, University of Michigan.

Crystal, D. (1987). *Cambridge Encyclopedia of the English Language*. Cambridge University Press.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fuchs, A. F. (1976). The neurophysiology of saccades. In Monty and Senders (Eds.), *Eye Movements and Psychological Processes*. Hillsdale, NJ: Lawrence Erlbaum.

Gray, W. D., John, B. E., and Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. *Human Computer Interaction*, 8:237–309.

Horswill, I. (In press). NLP on a mobile robot: Pipedreams and suggestions from active vision. Technical Report, American Association for Artificial Intelligence.

Jaffe, J. and Feldstein, S. (1970). *Rhythms of Dialogue*. New York: Academic Press.

John, B. E. (1988). *Contributions to Engineering Models of Human-Computer Interaction*. PhD thesis, Carnegie Mellon University.

John, B. E., Remington, R. W., and Steier, D. M. (1991). An analysis of space shuttle countdown activities: Preliminaries to a computational model of the NASA Test Director. Technical Report CMU-CS-91-138, Carnegie Mellon University Computer Science.

Laird, J. E., Newell, A., and Rosenbloom, P. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64.

Lehman, J. F., Lewis, R. L., and Newell, A. (1991). Natural language comprehension in Soar: Spring 1991. Technical Report CMU-CS-91-117, Carnegie Mellon University Computer Science.

Lewis, R. L. (1993). *An Architecturally-based Theory of Human Sentence Comprehension*. PhD thesis, Carnegie Mellon University. Also available as Technical Report CMU-CS-93-226.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction*, pages 353–388.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Norwine, A. C. and Murphy, O. J. (1938). Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, 17:281–291.

Rubinoff, R. and Lehman, J. F. (1994). Real-time natural language generation in nl-soar. In *7th International Workshop on Natural Language Generation, Kennebunkport, Maine*.

Torrance, M. C. (In press). Two case studies is active language use. Technical Report, American Association for Artificial Intelligence.

Traum, D. R., Allen, J. F., Ferguson, G., Heeman, P. A., Hwang, C. H., Kato, T., Martin, N., Poesio, M., and Schubert, L. K. (1994). Integrating natural language understanding and plan reasoning in the TRAINS-93 conversation system. Technical Report, American Association for Artificial Intelligence.

Westheimer, G. (1954). Mechanism of saccadic eye movements. *A.M.A. Archives of Opthalmology*, 52:710–723.

Wiesmeyer, M. D. (1992). *An Operator-Based Model of Human Covert Visual Attention*. PhD thesis, University of Michigan.