+

AUGMENTED TRANSITION NETWORKS (ATNs) FOR DIALOG CONTROL: A LONGITUDINAL STUDY

Curry Guinn
University of North Carolina Wilmington
Wilmington, NC 27403
Robert Hubal
RTI International
Research Triangle Park, NC 27709, USA
guinnc@uncw.edu, rhubal@rti.org

ABSTRACT
Our research team has implemented over a dozen spoken natural language dialog systems in varied domains over the past decade. Each system uses the same underlying dialog controller – an augmented transition network (ATN) – for maintaining a cohesive, natural conversation with the user. In this paper, we will examine the evolution of our use of ATNs, present statistical analysis of the features of our ATNs, and discuss lessons learned.

KEY WORDS
Dialog, natural language processing, mixed initiative, augmented transition networks, virtual humans.

1. Introduction

Since approximately 1996, our research team has worked on a series of PC-based applications in which the user interacts with responsive virtual characters. Applications have ranged from trauma patient assessment [1] to learning military tank maintenance diagnostic skills [2] to gaining skills in avoiding non-response during field interviews [3]. In these applications, the computer simulates a person's behavior in response to user input. Users interact with the virtual characters via voice, mouse, menu, and/or keyboard. We are certainly not alone in developing training, assessment, marketing, and other virtual human applications (see, e.g., [4,5,6,7,8,9,10,11]), but the breadth across domains and the consistency of the underlying architecture allows us to measure our systems' performance longitudinally.

We have developed a dialog system architecture that enables users to engage in unscripted conversations with virtual humans and see and hear their realistic responses [12]. As seen in Figure 1, among the components that underlie the architecture are a Language Processor and a Behavior Engine. The Language Processor accepts spoken input and maps this input to an underlying semantic representation, and then functions in reverse, mapping semantic representations to gestural and speech output. Our applications variously use spoken natural language interaction [2], text-based interaction, and menu-based interaction. The Behavior Engine maps Language Processor output and other environmental stimuli to virtual human behaviors. The underlying data structure of the Behavior Engine is an augmented transition network (ATN) to be described in more detail in Section 3. These behaviors include decision-making and problem solving, performing actions in the virtual world, and spoken dialog. The Behavior Engine also controls the dynamic loading of contexts and knowledge for use by the Language Processor. The virtual characters are rendered via a Visualization Engine that performs gesture, movement, and speech actions, through morphing of vertices of a 3D model and playing of key-framed animation files (largely based on motion capture data). Physical interaction with the virtual character (e.g., using medical instruments) is realized via object-based and instrument-specific selection maps [13]. These interactions are controlled by both the Behavior Engine and Visualization Engine.

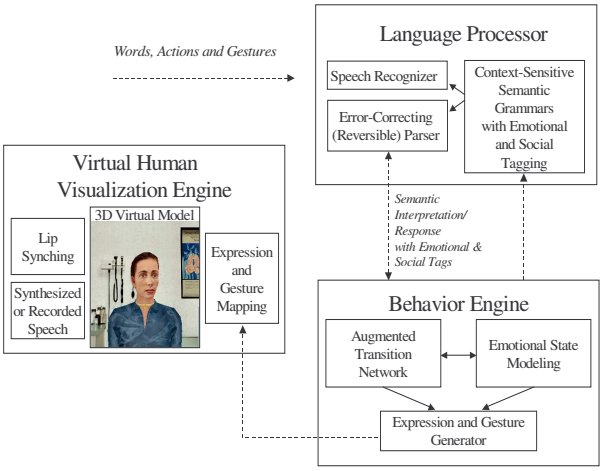


Figure 1: Dialog System Architecture

The architecture was designed to allow application creators flexibility in assigning general and domain-specific knowledge. Hence, our virtual humans discuss relevant concerns or excuses based on specific setup variables indicating knowledge level and initial emotional

state. Our personality models and emotion reasoning are based on well-accepted theories that guide realistic emotional behavior [14,15,16,12]. After user input, we update emotional state based on lexical, syntactic, and semantic analyses [17,18].

2. The Applications

Spoken dialog systems can be used in a wide variety of tasks. Our systems have focused on three areas: 1) question-answer kiosk applications, 2) "soft-skills" training applications, and 3) skill assessment.

2.1 Question/Answer Kiosk Applications

Question/Answer kiosks are commercial applications designed to enhance tradeshow booths and presentations by allowing customers to talk with a virtual tradeshow attendant that is able to answer queries about the company and its products (Figure 2a as example). In this paper, we will examine the ATN structure of three kiosks deployed at the Space Congress held in April 1999, 2) the American Society for Training and Development (ASTD) International Conference & Exposition held in May 1999, and 3) John Deere Exhibition in December 1999. Performance evaluations metrics may be found in [19] and, as a sample, include information such as number of separate conversations (335 users at Space Congress), average conversation length (61.4 seconds), and average number of dialog turns (5.6).

2.2 Training Applications

A number of training applications were developed to allow trainees to practice their "soft skills" – interpersonal communication skills in their performance of their jobs. In all of these training applications, the trainee converses using natural language with a virtual human who acts as client/subject to the trainee.

2.2.1 Bank Teller Trainer

For training bank tellers, we developed virtual characters that acted as bank customers with various transactions. Particular attention was placed on varying the customer's emotions and determining whether the teller trainee would react appropriately in dealing with a potentially difficult customer. A second virtual human acted as a coach as the situation warranted (Figure 2b) [20].

2.2.2 Survey (Door-to-Door) Trainer

In conjunction with RTI International's unit that conducts door-to-door surveys, we developed a trainer that allows surveyors to practice the difficult task of gaining interviewee compliance. Details of this implementation and evaluation can be found in [21].

2.2.3 Standardized Asthma Profile Trainer

Following some previous work on medical patient simulation [1], we developed a module that allowed a clinician to practice administering an asthma profile survey. In this system, the trainee must ask a series of questions and follow-up questions related to a patient's reported symptoms [22].

2.2.4 Survey (Telephone) Trainer

Similar to the door-to-door survey trainer, this system trains interviewers in gaining compliance over the telephone. In this system, there is no visible virtual human; however, the underlying architecture is identical for generating the virtual character's behavior and responses [23].

2.2.5 JUSTTALK Trainer

In this National Institute of Justice-funded project, police officers are trained at identifying mentally disturbed individuals and de-escalating confrontations (Figure 2d). Evaluation results indicate high usability (88% found the trainer easy to use) and effectiveness (59% found the simulation better or comparable to classroom discussion) based on interviews with law enforcement officers who participated in the training [18, 24].

2.2.6 Pediatric Clinician Trainer

Using the same technology we developed specific interactive training sessions using virtual pediatric characters to explore educational issues related to pediatric care (Figure 2e) [25]. Three scenarios were created – for this analysis we just examine one of them, a health assessment of a teenage girl while her father is present in the room.

2.2.7 World Trade Center Informed Consent Trainer

This survey training tool was developed for a specific study where interviewers obtained informed consent from participants involved in a mental, physical, and emotional health study on the short- and long-term effects of the World Trade Center bombing (Figure 2f) [26,27].

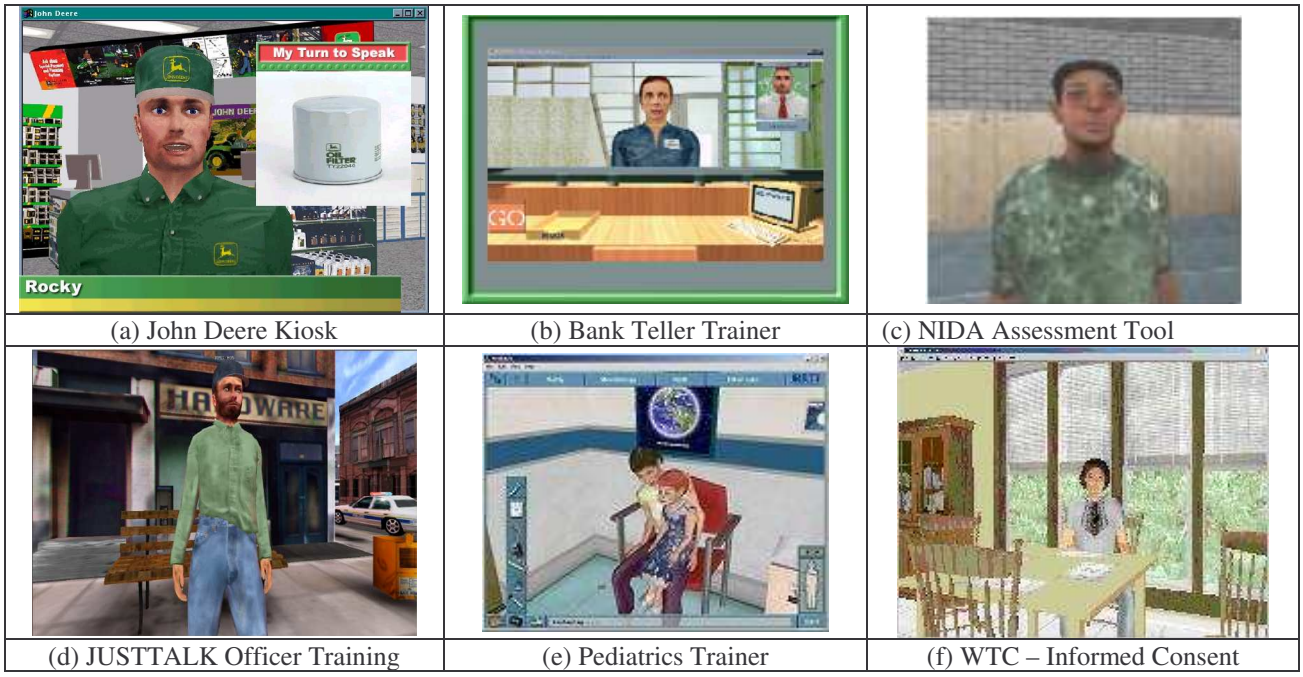
2.3 Profiler Application

A third class of applications has been developed to access user's behaviors in simulated environments.

2.3.1 NIDA Risk Assessment Profiler

Primarily, this National Institute on Drug Abuse-funded application has been used to study the behavior of adolescents who are labeled "at-risk" of engaging in risky behaviors and have undergone various interventions (Figure 2c). Results of these studies show that the system is accurate in judging which users have been categorized "at-risk" by experts [28]. Three different scenarios have been developed. Here we just present the statistics for two, one involving a scenario involving stolen goods and one involving smoking.

We also developed under National Institute of Justice funding a set of scenarios for assessing social competency skills of prisoners [29].



(a) John Deere Kiosk | (b) Bank Teller Trainer | (c) NIDA Assessment Tool

(d) JUSTTALK Officer Training | (e) Pediatrics Trainer | (f) WTC – Informed Consent

Figure 2: Spoken Dialog Applications with Virtual Humans

3. ATN Data Structure for Dialog

The Behavior Engine maps the output of the Language Processor and other environmental stimuli to virtual human behaviors. These behaviors include decision and problem solving, performing actions in the virtual world, changes in facial and body expression, and spoken dialog. The Behavior Engine also controls the dynamic loading of contexts and knowledge for use by the Language Processor. The Behavior Engine uses the semantic interpretation generated by the Language Processor to assist in determining the behavior of the virtual human. The underlying architecture of the Behavior Engine is an Augmented Transition Network (ATN). The structure of a node in the network includes fields for:

- *Name*. The name of the node,
- *Type*. The type of the node. Each virtual human in the simulation has its own node type (an "avatar" node) and there is one node type when an input from the user is expected (a "Normal" node). During execution, when the system reaches a Normal node, the system stops and waits for a user input which must occur before a timeout that is set within the network.
- *Grammar*. A grammar or set of grammars that should be loaded for speech recognition or language generation. This enables context-sensitive loading of grammars improving recognition and parsing accuracy.
- *Response*. The semantic information the virtual human would like to convey to the user.
- *VR Action*. Events that should occur in the virtual environment when this node is reached.
- *Action*. ATN variable assignments that should occur when this node is reached.
- *Transition*. The name of the node to make a transition to if the condition is satisfied in the Transition Condition field described below.
- *Transition Action*. Actions (variable assignments) that should occur if this transition is taken.
- *Transition Condition*. A Boolean expression using variable settings and inputs from the Language Processor.

```
wait_on_input
normal
Grammar: "intro.gram default.gram"
Response: "inform(offer_assistance)"
VRString: "raise(eyebrows)"
Action: "MENTOR = 0.8"
Transition: proc_command
TransitionAction: "INPUT = command"
Conditional: "command(CONTENT)"
Transition: proc_query
TransitionAction: "INPUT = query"
Conditional: "query(CONTENT)"
```

Figure 3: Example ATN Normal Node

The example Normal node named *wait_on_input* is defined in Figure 3 such that if this node is reached: 1) a spoken response matching the semantics *inform(offer_assistance)* will be uttered, 2) the character will raise its eyebrows, 3) the mentoring level is set to 0.8, 4) the grammars *intro* and *default* are loaded, 5) if the semantic input matches *command(X)* for some *X*, the system will transition to node *proc_command* and set *CONTENT* equal to *X*, or 6) if the semantic input matches *query(X)* for some *X*, the system will transition to node *proc_query* and set *CONTENT* equal to *X*.

Application	Semantics	Network Nodes	Normal (Speech) Nodes	Average Transitions Per Node	State Variables	Date
Space Congress (Kiosk)	38	107	2	4.401869	35	4/23/1999
Bank Teller	56	522	29	4.32567	98	5/13/1999
ASTD (Kiosk)	51	216	3	4.430556	36	5/26/1999
Door-to-door Survey	71	446	39	3.091928	119	10/21/1999
Deere (Kiosk)	34	171	1	4.140351	38	11/24/1999
Virtual Asthma Patient	52	355	20	2.309859	100	12/23/1999
Telephone Survey	23	109	2	4.229358	58	11/6/2001
NIDA StolenGoods	11	39	2	1.948718	56	11/20/2001
NIDA Smoking	15	41	2	1.878049	56	12/30/2002
JUSTTALK	81	284	2	4.169014	120	5/22/2003
Pediatrics Teen	97	586	2	2.1843	120	8/15/2003
WTC Informed Consent	66	227	2	4.911894	69	10/9/2003

Table 1: ATN Statistics for Each Dialog Application

Each transition is represented by the 3-tuple (Transition, Transition Action, and Transition Condition). There is normally more than one transition leading out of a node, though at least one must have a Transition Condition that defaults to true. If multiple Transition Conditions are satisfied at a particular node, then one is selected non-deterministically.

4. Study Results

Table 1 presents each application studied here, the number of nodes in that application's ATN structure, the number of semantic inputs possible from the user[1], the number of nodes in the network, the number of Normal (input) nodes, the average number of transitions per node, the number of variables used in the ATN, and the date of deployment[2].

4.1 Semantics versus network size

Intuitively, we would expect as the number of semantic inputs from the user increases, we should see a larger, more complex network. Our intuition is correct – the more options the user has as input, we see a roughly linear increase in the number of nodes in the ATN dialog structure as illustrated in Figure 4.

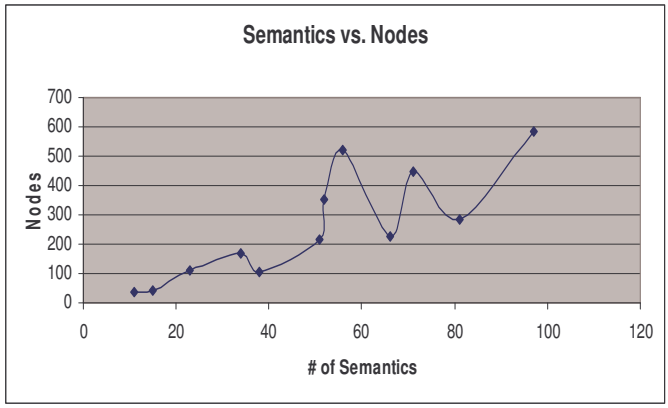


Figure 4: Semantic Complexity vs. Node Complexity

4.2 Variables used to control network size

This linear increase is encouraging. One might suspect that as the number of inputs increase the network size might grow exponentially. In practice, this does not seem to be the case. However, part of the reason the network does not increase exponentially is the use of variables within the ATN. If we were coding our dialog structure with a finite state machine (using no variables), we would find an exponential increase in nodes. However, each variable used has the potential to cut the number of nodes in half or more. For instance, suppose we want our virtual human to act either angry or happy. With a finite state machine, we would have to build two separate subnetworks to represent the behaviors for these two conditions. With an ATN, we can have variables keep track of the emotional state (ANGER = 0.5; HAPPINESS = -0.9) and use those variables in deciding which arc to take from a particular node.

[1] Note that the number of semantics is not the same as the number of possible inputs in a natural language processing system. The spoken or typed input will be mapped to a set of underlying semantic meanings. This is a many-to-one relationship.

[2] Work is still ongoing in some applications. The dates and networks shown here are those of initial deployment.

4.3 Longitudinal decrease in the number of Normal nodes

Training applications that were developed chronologically earlier (Banking, Door-to-door survey, Asthma) tend to have many more Normal (Speech) nodes that have the effect of increasing the total number of nodes and network complexity. By contrast, later training applications had many fewer Normal nodes as illustrated in Figure 5. One of the motivations for having different Normal nodes was to load different sets of grammars for speech recognition and parsing depending on context. By creating these different nodes (which share many of the same inputs and transitions), the knowledge engineering and network maintenance tasks became much more difficult. Networks were being replicated unnecessarily. It was found in practice that having one or two nodes that handle the user input was far preferable. The dynamic loading of different grammars is accomplished by setting a GRAMMAR variable rather than creating new nodes.

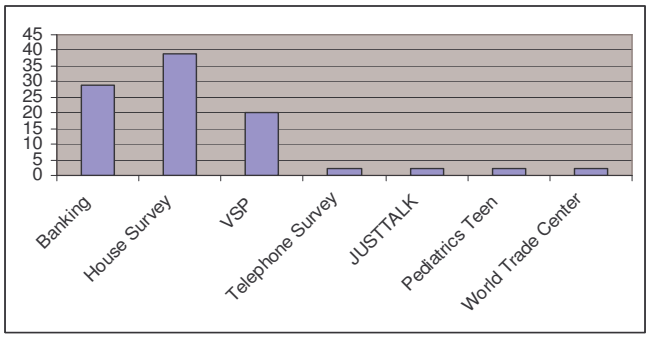


Figure 5: Normal Nodes per Training Application in Chronological Order

4.4 Increase in number of total nodes versus network branching complexity

The most mature and complex application is the set of Pediatric training scenarios. This application has the highest number of semantic inputs (97) and the greatest number of nodes (586) – over twice as many nodes as the slightly less semantically complex JUSTTALK (284 nodes). However, the Pediatric application has strikingly lower inter-node complexity as indicated by the average number of transitions per node (2.18) compared to the other training applications like JUSTTALK (4.17) or Telephone Survey (4.23). A network with less node inter-connectivity is easier to trace and debug. At each individual node, there are fewer options to evaluate to determine where to go next. From a maintenance and reliability standpoint, increasing the number of nodes in order to create a larger but less-interconnected network is preferable.

5. Study Results

A trend over time indicates that our training applications are increasing in semantic complexity. As speech recognition improves, the demands on conversational systems will become even greater. Having a linear increase in the number of nodes is beneficial for reducing costs in development and reliability in operation. To do so, we found it necessary to increasingly use variables to keep track of dialog state. This use of variables does have a negative side effect – it reduces the ease of reading and tracing networks – therefore, it is a tradeoff. In the most semantically complex application, we even chose to increase the number of nodes rather than increase the complexity of the network inter-connections precisely because we needed more reliable development and deployment. Further work will quantify measurements related to cost, development, and maintenance of dialog networks.

Acknowledgements

The development and studies described here were performed under awards #290-00-0021 from the Agency for Healthcare Research and Quality, #1-S07-RR18257-01 from the National Institutes of Health, NIDA award #5-R01-DA14813-02, 2000-RD-CXK002, National Institute of Justice, and #EIA-0121211 from the National Science Foundation. Points of view in this document are those of the authors, and do not necessarily represent the official position of any of the above-listed agencies.

References

1. Kizakevich, P.N., McCartney, M.L., Nissman, D.B., Starko, K., & Smith, N.T. (1998). Virtual Medical Trainer: Patient Assessment and Trauma Care Simulator. In J.D. Westwood, H.M. Hoffman, D. Stredney, & S.J. Weghorst (Eds.), *Art, Science, Technology: Healthcare Revolution*. Amsterdam, Holland: IOS Press, 309-315.

2. Guinn, C.I., & Montoya, R.J. (1998). Natural Language Processing in Virtual Reality. *Modern Simulation & Training*, 6, 44-45.

3. Camburn, D.P., Gunther-Mohr, C., & Lessler, J.T. (1999). Developing New Models of Interviewer Training. *Proceedings of the International Conference on Survey Nonresponse*. Portland, OR.

4. André, E., Rist, T., & Müller, J. (1999). Employing AI Methods to Control the Behavior of Animated Interface Agents. *International Journal of Applied Artificial Intelligence*, 13 (4-5), 415-448.

5. Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz Studies – Why and How. *Knowledge-based Systems*, 6(4), 258-266.

6. Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the Tutoring Research Group (2000). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1, 35-51.

7. Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., & Bhogal, R. (1997). The Persona Effect: Affective Impact of Animated Pedagogical Agents. *Proceedings of the Human Factors in Computing*

Systems Conference, (pp. 359-366). New York, NY: ACM Press.

8. Lundeberg, M., & Beskow, J. (1999). Developing a 3D-Agent for the August Dialogue System. *Proceedings of the Auditory-Visual Speech Processing Conference*. Santa Cruz, CA.
9. Olsen, D.E. (2001). The Simulation of a Human for Interpersonal Skill Training. *Proceedings of the Office of National Drug Control Policy International Technology Symposium*. San Diego, CA.
10. Rickel, J., & Johnson, W.L. (1999). Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control. *Applied Artificial Intelligence*, 13, 343-382.
11. Rousseau, D., & Hayes-Roth, B. (1997). Improvisational Synthetic Actors with Flexible Personalities. *KSL Report #97-10*, Stanford University.
12. Guinn, C., Hubal, R., Frank, G., Schwetzke, H., Zimmer, J., Backus, S., Deterding, R. Link, M. Armsby, P., Caspar, R., Flicker, L., Visscher, W., Meehan, A., and Zelon, H. (2004). Usability and Acceptability Studies of Conversational Virtual Human Technology, *5th SIGdial Workshop on Discourse and Dialogue*, Boston, MA.
13. Zimmer, J., Kizakevich, P., Heneghan, J., Schwetzke, H., & Duncan, S. (2003). The Technology Behind Full Body 3D Patients. Poster presented at the *Medicine Meets Virtual Reality Conference*. Newport Beach, CA. http://www.rvht.info/publications.cfm.
14. André, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (2000). Exploiting Models of Personality and Emotions to Control the Behavior of Animated Interface Agents. *Proceedings of the International Conference on Autonomous Agents* (pp. 3-7). Barcelona, Spain.
15. Ortony, A., Clore, G.L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge, England: Cambridge University Press.
16. Russell, J.A. (1997). How Shall an Emotion Be Called? In R. Plutchik & H.R. Conte (Eds.), *Circumplex Models of Personality and Emotions* pp. 205-220. Washington, DC: American Psychological Association.
17. Guinn, C.I., & Hubal, R.C. (2003) Extracting Emotional Information from the Text of Spoken Dialog, *9th International Conference on User Modeling*, Johnstown, PA.
18. Hubal, R.C., Frank, G.A., & Guinn, C.I. (2003). Lessons Learned in Modeling Schizophrenic and Depressed Responsive Virtual Humans for Training. Proceedings of the Intelligent User Interface Conference. Miami, FL
19. Guinn, C.I., & Hubal, R.C. (2004). An Evaluation of Virtual Human Technology in Informational Kiosks, *Proceedings of International Conference on Multimodal Interfaces* (ICMI '04), State College, PA.
20. Hubal, R.C., & Guinn, C.I. (2002). A Mixed-Initiative Intelligent Tutoring Agent for Interaction Training. *Intelligent User Interface Conference*.
21. Camburn, D.P., Gunther-Mohr, C., and Lessler, J.T. (1999) Developing New Models of Interviewer Training. *International Conference on Survey Nonresponse*, Portland OR.
22. Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., & West, S.L. (2000). The Virtual Standardized Patient--Simulated Patient-Practitioner Dialogue for Patient Interview Training. In J.D. Westwood, H.M. Hoffman, G.T. Mogel, R.A. Robb, & D. Stredney (Eds.), *Envisioning Healing: Interactive Technology and the Patient-Practitioner Dialogue*. IOS Press: Amsterdam, 133-138.
23. Link, M., Armsby, P., Hubal, R. and Guinn, C. (2006). Accessibility and acceptance of responsive virtual human technology as a survey interviewer training tool, *Computers in Human Beha*vior 22(3):412-426.
24. Frank, G.A., Guinn, C.I., Hubal, R.C., Stanford, M.A., Pope, P., & Lamm-Weisel, D. (2002). JUST-TALK: An Application of Responsive Virtual Human Technology. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*.
25. Deterding, R., Milliron, C., and Hubal, R. (2005). The Virtual Pediatric Standardized Patient Application: Formative Evaluation Findings. Studies in Health Technology and Informatics, (111):105-107.
26. Hubal, R., Guinn, C, Visscher, W., Studer, E., and Sparrow, D. A. (2004). Synthetic Character Application for Informed Consent, *AAAI Fall Symposium on Dialogue Systems for Health Communication*, Washington, DC
27. Hubal, R., & Day, R., (*in press*) Informed consent procedures: An experimental test using a virtual character in a dialog systems training application. *J. of Biomedical Informatics.*
28. Paschall, M., Fishbein, D., Hubal, R, and Eldreth, D. (2005). Psychometric properties of virtual reality vignette performance measures: a novel approach for assessing adolescents' social competency skills. *Health Education Research*, v. 20(1).
29. Fishbein, D., Scott, M, Hyde, C., Newlin, D., Hubal, R., Serin, R., Chrousos, G., & Alesci, S. (2006). Neuropsychological and Emotional Deficits Predict Correctional Treatment Response. Final Report, Submitted to the Office of Justice Program, National Institute of Justice, Award #2002-MU-BX-0013.