DESIGN AND EVALUATION OF CONVERSATIONAL MODELING METHODS FOR INTERPERSONAL SIMULATION

By

BRENT H. ROSSEN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2011

To my Wife, Mom, and Dad for their love and support

ACKNOWLEDGMENTS

I thank my research advisor and supervisory committee chair, Dr. Benjamin Lok. He taught me how to identify a problem, conduct research, and, most importantly, how to communicate our findings with the rest of the world. None of this work would have been achievable without his advice, direction, and support. I also thank my supervisory committee members, Dr. Paul Fishwick, Dr. Douglas Dankel, Dr. Mike Robinson, and Dr. Juan Cendan for their ideas and support in this research.

I thank my collaborators in healthcare education for their work and insight: Dr. Carole Kimberlin and Dr. Diane Beck for their participation in the pharmacy study. I thank Dr. Scott Lind, Dr. Adriana Foster, Dr. Adeline Deladisma, Dr. Hevil Shah, Dr. Michael Crary, and Dr. Juan Cendan for their work on virtual patients and their invaluable feedback on Virtual People Factory.

Thank you to my collaborators in the Virtual Experiences Research Group: Dr. Aaron Kotranza, Dr. Kyle Johnsen, Dr. Andrew Raij, Dr. Regis Kopper, Joon Chauh, and Shivashankar Halan. The previous and ongoing work of these research collaborators inspired the methods and systems described in this dissertation. Further, their friendship, advice, and commiseration made the process of completing this research possible and often even fun.

Lastly, I thank my family. I thank my parents, Jan and Joel, for raising me to believe I can take on anything. And I thank my wife, Elizabeth, for her support through all the bumps and turns of my Ph.D. journey.

This research was funded by a University of Florida Alumni Fellowship and National Science Foundation Grant (IIS-0643557).

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

VH	Virtual human. A computer simulation of a human conversational partner.
CCM	Centralized Conversational Modeling. A typical approach of conversational modeling which follows the patterns established in the field of expert systems, in which a knowledge engineer acquires domain specific knowledge, and then translates that knowledge into a machine-readable format (section 1.1).
HDCM	Human-centered Distributed Conversational Modeling. A novel human-centered approach to conversational modeling used to facilitate natural language conversation (section 1.3.1, Chapter 3).
VHT	Virtual Human Templates. A novel method of Conversational Knowledge Reuse for conversational modeling based on the knowledge acquired for previously created similar conversational models (section 1.3.2, section 4.2).
DKS	Dynamic Knowledge Sharing. A novel method of Conversational Knowledge Reuse for conversational modeling based on the knowledge acquired for conversational models that are still undergoing development (section 1.3.3, section 4.3).
VPF	Virtual People Factory. A web-based application that implements the conversational modeling methodologies described in this dissertation (section 3.2).

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

DESIGN AND EVALUATION OF CONVERSATIONAL MODELING METHODS FOR INTERPERSONAL SIMULATION

By

Brent H. Rossen

December 2011

Chair: Benjamin Lok
Major: Computer Engineering

Interpersonal skills practice using virtual humans offers structured learning of interview skills, provides standardization of practice, and can facilitate learning about unusual conditions. However, the creation of virtual humans with the ability to understand and respond to natural language requires costly engineering by conversation knowledge engineers (generally computer scientists), and incurs logistical cost for acquiring domain knowledge from domain experts (generally educators). We address these problems using two novel methods entitled Human-centered Distributed Conversational Modeling and Conversational Knowledge Reuse. These methods facilitate collaborative development of virtual humans by two groups of end-users: domain experts (educators) and domain novices (students). We implemented these methods in a web-based authoring tool called Virtual People Factory. Using Virtual People Factory, medical and pharmacy educators create natural language virtual patient interactions. This dissertation presents the theoretical background for Human-centered Distributed Conversational Modeling and Conversational Knowledge Reuse; the implementation of the Virtual People Factory authoring tools; and studies showing that these methods have reduced the logistical cost of acquiring knowledge and thereby

improved the feasibility of applying virtual human based interpersonal simulation to real-world education.

CHAPTER 1
INTRODUCTION

Virtual humans (VHs) for natural language conversation are becoming increasingly popular for communication skills training. A VH is a computer simulation of a human conversational partner. VHs can communicate using speech, gestures, and facial expressions. VHs for interpersonal skills training have been shown to be useful for providing structured learning of interview skills and decreasing anxiety in potentially uncomfortable interviews (Deladisma et al., 2007b; Johnsen, 2008; Raij, 2009). They are also predicted to be useful for providing diverse interview experiences (Johnsen, 2008; Rossen et al., 2008). Projects in medicine, psychology, and the military have been created with collaborative effort between end-users (educators and students), artists, programmers, and conversation knowledge engineers (generally computer scientists) (Dickerson et al., 2005; Johnsen, 2008; Kenny et al., 2007; Kenny et al., 2008; Kotranza, 2009; Raij, 2009; Villaume et al., 2006). However, these collaborators find acquiring the knowledge necessary to simulate these conversations, and processing that knowledge into a conversational model, to be logistically difficult and time consuming (Dickerson et al., 2005; Glass et al., 2005; Kenny et al., 2007; Kenny et al., 2008; Villaume et al., 2006).

A conversational model is a digital representation of the knowledge necessary to conduct a conversation. Conversational models are used in VH simulations to recognize and respond to speech inputs. Preparing a VH to conduct a free-form conversation usually takes hundreds of hours over several months. For example, Vic is a VH created to play the role of a patient having stomach pain. Vic is capable of ten minutes of free-form conversation about his symptoms with a pharmacy student. Even for ten minutes

of conversation, a team of knowledge engineers and domain experts spent approximately 200 hours and over 6 months to obtain 75% accuracy in Vic's conversational simulation (Chapter 3). The labor-intensive nature of conversational model development restricts the number of VHs that can be created. As a result of such impractical time and expertise requirements, educators have been unable to create the number of scenarios needed to implement an interpersonal skills training curriculum, and are even further from creating a diverse training curriculum (Plant et al., 2005; Triola et al., 2007).

To address these problems, we propose two new methods based on human-centered computing (Sebe, 2010). The first method, Human-centered Distributed Conversational Modeling (HDCM) is a novel crowdsourcing method for developing conversational models. The second method, Conversational Knowledge Reuse is a novel method for reusing portions of previously acquired conversational knowledge to make new conversational models. Both of these methods center around minimizing the human effort involved in the process of collecting conversational knowledge and assembling that knowledge into a conversational model.

Applying HDCM and Conversational Knowledge Reuse to the problem of conversational modeling improves the efficiency of human effort used to model a conversation. This efficiency results in a significantly shorter time to produce a more accurate conversational model than previous methods. This reduction in time and effort increases the feasibility of implementing interpersonal skills training curricula using VHs.

The methods described in this dissertation provide insights that will open new areas of VH applications and make research possible that would have been difficult or

impossible to implement using previous methods. Thus, in developing and evaluating these methods for constructing VH conversational models, this dissertation provides innovations to both the study of VH user interfaces and their practical application in interpersonal simulation.

1.1 Problem Statement

The proposed methods address the modeling of natural language conversations for interpersonal skills education using VHs. Simulating a natural language conversation requires a robust conversational model that can recognize and respond accurately to a wide range of inputs. To develop a robust conversational model, knowledge engineers must acquire a conversation-specific corpus reflecting what the users will say to a VH (stimulus) and what the VH will say back (response) (Dickerson et al., 2005; Kenny et al., 2007; Kenny et al., 2008; Leuski et al., 2006; Reiter et al., 2003; Ruttkay et al., 2004). The stimuli-response space of these conversations is large, and often consists of thousands of stimuli linked to hundreds of responses.

Typical conversational modeling methods follow the patterns established in the field of expert systems, in which a knowledge engineer acquires domain specific knowledge, and then translates that knowledge into a machine-readable format (Shortliffe, 1976). Knowledge acquisition and engineering is cited as the biggest barrier in expert systems and natural language systems development (Reiter et al., 2003; Wagner, 1990). The time required for knowledge acquisition and translation limits the utility of knowledge engineer based methods for conversational modeling.

To develop natural language conversational models, knowledge engineers acquired information from the following resources: 1) recordings of people in natural or staged interactions; 2) Wizard of Oz (human-controlled) VH interactions; and 3) expert

descriptions of real-world interactions (Ruttkay et al., 2004). The conversational knowledge engineer uses these resources to seed the conversational modeling process by gathering the starting stimuli and translating those stimuli into a machine-readable conversational corpus. The knowledge engineer then refines the knowledge corpus by collecting knowledge from VH interactions with users, reviewing these user interactions, making changes to the corpus based on conversation errors, validating the corpus changes by consulting experts, and then repeating the process (Dickerson et al., 2005; Kenny et al., 2007; Kenny et al., 2008; Leuski et al., 2006).

We will hereafter refer to the described method as Centralized Conversational Modeling (CCM) because of the knowledge engineer's role as the hub for transferring information from experts and novices to the conversation corpus (conversational knowledge database) as shown in Figure 1-1. Because all of the knowledge is flowing through the knowledge engineer, the CCM process is slow, and experts report that communication bottlenecks are often frustrating (Dickerson et al., 2005; Glass et al., 2005; Kenny et al., 2007; Kenny et al., 2008; Villaume et al., 2006). Additional details on the challenges presented in this problem statement are provided in section 2.2.2.

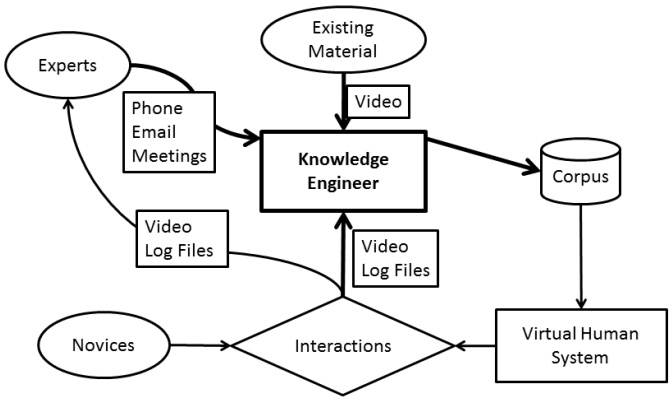

Figure 1-1. The flow of data in Centralized Conversational Modeling. All of the conversational data is piped through the knowledge engineer and into the corpus, causing a bottleneck.

1.2 Research Questions

The applicability of VH interpersonal simulation to real-world purposes is limited by the inefficiency of the process described in the problem statement. Based on this overview of the problem, we define three broad research questions, which we address through the research presented in this dissertation:

Research Question 1: What are the bottlenecks involved in the creation and use of virtual humans in interpersonal skills training?

This question is the starting point for understanding the problems we are confronted with, the current challenges in generating virtual humans.

Analyses of the existing material on the creation of VH conversational models were performed. These analyses provided insights into the challenges of generating VHs, the bottlenecks that slow down the conversational modeling process.

Research Question 2: What conversational modeling methods can we design to alleviate the problems caused by the bottlenecks identified in question 1?

The answers to this question will provide designers and researchers with new methods for modeling virtual human conversations.

The two methods presented are novel approaches to generating VHs. These methods are designed to improve the efficiency of conversational modeling, and thereby reduce the time required to create robust conversational models.

Research Question 3: Does use of the software, implemented based on the methods designed in response to question 2, result in efficient creation of robust conversational models?

This question aims to validate the conversational modeling methods by showing that they are implementable and that robust conversational models can be created more efficiently than previous methods.

Following the proposed methodologies, we implemented systems that follow each method. These systems are shown to facilitate the efficient creation of VH conversational models that are robust to a wide range of inputs. By addressing this question, we provide evidence that the methods succeeded in providing direction for the design of efficient conversational modeling software for use in a real-world domain.

1.3 Overview of Approach

The research into new conversational modeling methods was carried out through an iterative process of analysis, design, implementation, and evaluation (research questions 1-3). We analyzed the current state of the art of VH conversational modeling to determine bottlenecks in the process of the creation of diverse sets of VHs (research question 1 - Chapter 2). From this first analysis, the problem statement was derived (section 2.2).

We addressed the issues presented in the problem statement by applying lessons from human-centered design (Sebe, 2010), crowdsourcing (Howe, 2006; Singh et al., 2002) and human-computation (von Ahn and Dabbish, 2004). The combination of human-centered design, crowdsourcing, and human-computation offers inspiration for novel solutions to acquiring knowledge and processing that knowledge into VH conversational models. Based on these concepts, this dissertation presents two new methods for the efficient creation of VH conversational models, HDCM and Conversational Knowledge Reuse (research question 2 – Chapter 3, 4, 5).

We implemented these methods in a web-based application called Virtual People Factory. Virtual People Factory was evaluated for use in interpersonal skills training in the healthcare field (Filichia et al., 2010; Foster et al., 2010a; Foster et al., 2010b; Palathinkal, 2011; Peden et al., 2011; Pileggi and Childs, 2011; Rossen et al., 2010; Rossen et al., 2012; Rossen et al., 2009; Rossen and Lok, 2012; Shah et al., 2008; Shah et al., 2012; Shah et al., 2009a; Shah et al., 2009b; Surkunalingam et al., 2009). This dissertation presents three evaluations of the conversational modeling methods implemented in Virtual People Factory as well as a meta-analysis of the HDCM method used in real-world education (research question 3 – Chapter 3, 4, 5).

1.3.1 Human-centered Distributed Conversational Modeling

The Virtual People Factory implementation of the HDCM method was evaluated first by the researchers in collaboration with healthcare professionals (Rossen et al., 2009), and then independently by additional healthcare professionals to assess the system within real-world contexts (Foster et al., 2010a; Foster et al., 2010b; Peden et al., 2011; Pileggi and Childs, 2011; Shah et al., 2008; Shah et al., 2012; Shah et al., 2009a; Shah et al., 2009b; Surkunalingam et al., 2009). These evaluations were analyzed, and the advantages and limitations for the HDCM process were determined (research questions 1-3 - Chapter 3) (Rossen and Lok, 2012).

In contrast to the CCM method described in the problem statement, the HDCM method proposes that VH users (as opposed to knowledge engineers) generate the model themselves through a crowdsourcing process. Using HDCM, domain experts and novices collaborate to teach the VH how to converse. HDCM results in a flow of data, Figure 1-2, that does not have the knowledge engineer bottleneck as in CCM, Figure 1-1. Domain novices speak with the VH, which gathers new stimuli, and the domain experts add new responses to these stimuli. The VH system itself processes the knowledge, which eliminates the CCM bottleneck of piping information through the knowledge engineer. Effectively, the novices and the expert collaborate to teach the VH how to conduct this domain-specific conversation – this process is both human-centered and embodies the ideas of crowdsourcing with human-computation.

HDCM applies the ideas of crowdsourcing (Howe, 2006; Singh et al., 2002) and human-computation (von Ahn and Dabbish, 2004) to the problem of enumerating the stimuli-response space of a conversation. Our evaluation results demonstrate that HDCM shortens the time to model the conversation, and the resulting VH

conversational model is more comprehensive. Further, the domain experts drive the HDCM process, thus allowing the experts to direct the creation of VH-based training curricula and to focus the material on learning goals (Chapter 3).

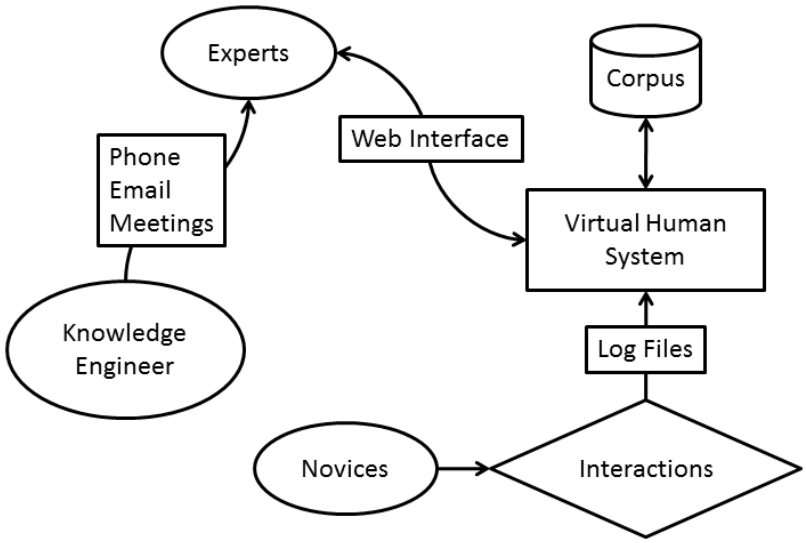

Figure 1-2. The flow of data in Human-centered Distributed Conversational Modeling. Conversational data flows from Experts and Novices into the Virtual Human System, which processes information and stores it in the corpus.

There are, however, limitations to the HDCM process including expert time requirements and end-user (novice) availability (Rossen and Lok, 2012). While the time is shorter than the amount previously required, many experts still perceive the process to be difficult (Chapter 3). HDCM also requires a large group of end-users. Further, these users must interact with VHs who do not yet have a response to many stimuli, meaning the conversational models will have a high error rate. These in-development VHs may provide limited educational value for novices (students) helping to create the conversational model. We were therefore motivated to make maximal use of the knowledge acquired from novice interactions; this motivation inspired the development of the Conversational Knowledge Reuse methods.

1.3.2 Conversational Knowledge Reuse: Virtual Human Templates and Dynamic Knowledge Sharing

To alleviate the identified issues in HDCM and to expand on the applicability of HDCM, we designed, implemented, and evaluated two additional conversational modeling methods based on the concept of Conversational Knowledge Reuse. Conversational Knowledge Reuse is the idea that we may leverage knowledge acquired for one VH in the creation of a different VH. The methods implementing Conversational Knowledge Reuse are *Virtual Human Templates* and *Dynamic Knowledge Sharing*. These new methods were implemented within the Virtual People Factory application. We then re-evaluated the Virtual People Factory application within the context of a healthcare course to examine if students could create VHs themselves, a task previously found to be infeasible (research questions 1-3 - Chapter 4) (Villaume et al., 2006).

Conversational Knowledge Reuse alleviates the limitations of HDCM and expands the applicability of VHs. The concept of Conversational Knowledge Reuse was based on the observation of significant overlap in knowledge between similar conversational models. Much of the knowledge acquired in the initial stages of HDCM consists of basic questions that are common to the interaction domain (medicine/psychology/military). These initial “seeding” stages cover basic utterances used in conversations; typical questions include “What is your name?” and “How old are you?” Repeatedly acquiring this knowledge for each conversational model takes up the time of both novice and expert users, thereby using up two of our limited resources.

Conversational Knowledge Reuse allows us to leverage this overlap of knowledge to seed new conversational models based on previous conversational models. For the

Virtual Human Templates method, knowledge engineers extract information from existing conversational models and generalize that knowledge into a Virtual Human Template. The Virtual Human Template is then customized by a VH author and that knowledge is used to generate a new seeded conversational model (Chapter 4). Virtual human templates streamline the process of seeding the conversational model. The initial seeding allows the subsequent use of HDCM to gather stimuli that are more specific to the current scenario, thereby using novice time more efficiently.

A second method of leveraging the similarity of information in conversational models is dynamically sharing data between closely related conversational models. When similar conversational models are created at the same time, we can improve these models by making sure that each of the conversational models can answer questions asked of any of the conversational models. This method is called *Dynamic Knowledge Sharing* and is embedded in the Virtual People Factory system.

During the conversational modeling process, the Virtual People Factory system automatically gathers questions asked of other similar conversational models. The system then examines if each conversational model contains each question, and shares questions the conversational model does not currently know. By sharing knowledge between the similar conversational models, the models can be created more rapidly and cover a greater portion of potential inputs. This again allows the process to further leverage the knowledge provided from novice users.

The advantages of *Virtual Human Templates* and *Dynamic Knowledge Sharing* further reduced the time requirements for enumerating the stimulus-response space of a conversation. In Chapter 4, we explore the use of these methods in the context of a

healthcare course, and examine if by using these methods healthcare students can rapidly generate robust VH patients, a task that was previously infeasible due to time requirements. The report on this evaluation presents a detailed analysis regarding the sources of conversational knowledge, the efficiency of the conversational modeling process, and the quality of the resulting conversational models.

1.3.3 Conversational Knowledge Reuse: Virtual Human Bootstrapping

We present an additional Conversational Knowledge Reuse method for reusing acquired knowledge to bootstrap the creation of new VHs, known as *Virtual Human Bootstrapping*. *Virtual Human Bootstrapping* allows authors to reverse the roles of VH interactions. The previously described VHs have a fixed role in a conversation. Medical students practice interviewing by conversing with a VH patient (Johnsen et al., 2005); teachers re-enact classroom situations with virtual students (Dieker et al., 2007); and soldiers learn conflict resolution in interactions with virtual civilians (Hill et al., 2003). During development of these human-VH interactions, the role played by the human and the role played by the VH are fixed. In the examples given, the soldier, teacher, and doctor are humans, while the civilian, student(s), and patient are VHs. We have observed that reversing these roles (e.g. a human patient with a VH doctor) will facilitate interpersonal skills training for additional large populations (Rossen et al., 2010). However, traditional methods of creating a VH for each role would double development time and effort. With *Virtual Human Bootstrapping*, authors can rapidly reverse the roles in these conversations, thereby allowing VHs to play the previously human side of the interaction (research questions 1-3 - Chapter 5).

We implemented the *Virtual Human Bootstrapping* method in an application called The Roleplay Trainer Creator. The Roleplay Trainer Creator generates VHs that are

compatible with the Virtual People Factory system. In Chapter 5, we describe a study which used The Roleplay Trainer Creator to generate a VH medical student from the conversational knowledge of a VH patient. This VH medical student was then used for training standardized patient actors to conduct a practice medical interview (Rossen et al., 2010). To the best of our knowledge, training standardized patients using VHs is a previously unexplored area of VH applications, and this evaluation demonstrates the feasibility of applying VHs to this new application area.

1.4 Thesis

We introduce the concepts of HDCM and Conversational Knowledge Reuse for conversational modeling.

> Thesis Statement: *The proposed conversational modeling methodologies improve the efficiency of the human effort used to model a conversation. Improved efficiency results in a significantly shorter time to produce more accurate conversational models than previous methods. This reduction in time and effort enhances the applicability of virtual humans to real-world interpersonal skills education.*

1.5 Innovations

This dissertation provides innovations in the areas of Human Computer Interaction in the subfield of Interpersonal Simulation and Artificial Intelligence in the subfield of Conversational Modeling. The innovations include the conversational modeling methodologies, novel software implementations, and evaluations of their efficacy.

Conversational Modeling Methods. We designed the methods of Human-centered Distributed Conversational Modeling (Chapter 3) and Conversational Knowledge Reuse (Chapter 4, 5). The Conversational Knowledge Reuse concept is embodied in three methods: Virtual Human Templates, Dynamic Knowledge Sharing,

and Virtual Human Bootstrapping. These methods were designed based on analyses of the current state of conversational modeling for virtual humans.

Software. We implemented the conversational modeling methods in software for creating VH conversational models for interpersonal skills training. These methods were implemented within Virtual People Factory (Chapter 3, 4), Virtual Human Templates (Chapter 4), and The Roleplay Trainer Creator (Chapter 5). These implementations have been released for research use and have been shown to be effective for real-world educational training applications (Filichia et al., 2010; Foster et al., 2010a; Foster et al., 2010b; Halan et al., 2010; Halan et al., 2012; Jackson, 2010; Palathinkal, 2011; Shah et al., 2008; Shah et al., 2009a; Shah et al., 2009b; Surkunalingam et al., 2009).

Evaluations. The Human-centered Distributed Conversational Modeling method was evaluated first in collaboration with a single expert, and then a meta-analysis was conducted on the method used in real-world practice (Rossen et al., 2009; Rossen and Lok, 2012). The Conversational Knowledge Reuse methods of Virtual Human Templates and Dynamic Knowledge Sharing were evaluated within the context of a healthcare training course wherein student participants created their own VH conversational models and used those models for practicing their interview skills (Rossen et al., 2012). The Virtual Human Bootstrapping method was evaluated by an expert participant, and a pilot study was performed investigating the applicability of the generated VHs to training (Rossen et al., 2010). The methods and software were evaluated in multiple real-world scenarios and found to be efficient for the creation of robust VH conversational models for interpersonal skills training (Chapter 3, 4, and 5).

CHAPTER 2
REVIEW OF LITERATURE

2.1 Interpersonal Simulation: Virtual Human Training Applications

Researchers have developed VH interpersonal simulation to provide the advantages of standardization, availability, immediate feedback, and diversity. Compared to human actors, VH conversational partners provide a greater degree of standardization and more consistent learner experiences (Johnsen, 2008; Kenny et al., 2008). VH partners are often autonomous, and so can provide anytime-anywhere training (Johnsen, 2008; Rizzo et al., 2010). Through after-action reviews, VH interactions can provide immediate feedback, and detailed visualizations can enhance learning experiences (Raij, 2009). VH interpersonal simulation can also provide diverse experiences such as abnormal findings (cranial nerve damage, a breast mass, or a facial burn) (Kotranza, 2009), ethnically diverse visual representations (Rossen et al., 2008), and using the work described in this dissertation, they can provide a greater variety of conditions and experiences (Rossen and Lok, 2012).

Because of these advantages, researchers have begun adopting VH interpersonal simulation in fields where interpersonal skills are essential to job performance. These fields include military leadership (Deaton et al., 2005; Hill et al., 2003; Kenny et al., 2007; Rizzo et al., 2010; Traum, 2008), mental health assessment (Kenny et al., 2008; Rizzo et al., 2010), and medical interviewing (Hubal and Day, 2006; Manganas et al., 2004; Stevens et al., 2005).

In military simulations, users learn tactics and negotiation strategies by interacting with VHs civilians, combatants, and fellow team members (Deaton et al., 2005; Hill et al., 2003; Kenny et al., 2007; Rizzo et al., 2010; Traum, 2008). In the field of mental

health care, clinical psychology students practice clinical interview skills and learn strategies to motivate patients to seek further counseling (Foster et al., 2010a; Foster et al., 2010b; Kenny et al., 2008; Parsons et al., 2008; Rizzo et al., 2010). In medicine, medical students learn patient interviewing skills with virtual patients (Hubal et al., 2000; Johnsen, 2008; Kotranza, 2009; Raij, 2009).

The collective goals of these research efforts are to examine the efficacy of learning experiences (Deladisma et al., 2007a; Parsons et al., 2008; Shah et al., 2008), to examine the feasibility of modeling human behavior with VHs (Dickerson et al., 2005; Kenny et al., 2008), to evaluate user behavior during and after interactions (Deladisma et al., 2007b; Iacobelli and Cassell, 2007; Zanbaka et al., 2007), and to construct cognitively accurate dialogue modeling architectures (Ellaway and McGee, 2008; Traum, 2008). Through that research, effective interaction has progressed, but speed and independent function of domain experts remains problematic (Huang et al., 2007; Triola et al., 2007). The solutions presented in this dissertation focus on making these VH technologies practical. With rapid and accurate knowledge acquisition, we increase the potential for widespread adoption of VHs as training tools.

2.1.1 Healthcare Interview Training

The conversational modeling methods proposed in this dissertation were developed within the field of healthcare interview training. Healthcare educators use interview training to help students develop their patient interviewing skills. Patient interviewing skills are crucial in all areas of healthcare. Interpersonal skills training results in both improved patient care and reduced lawsuits (Jenkins and Fallowfield, 2002; Vincent et al., 1995). Healthcare students traditionally train for these interviews by interacting with real patients in various healthcare environments and shadowing clinical

practitioners (Barrows, 1987). Given the importance of accurate diagnoses resulting from patient interviews, there is demand for additional patient interview practice (Itin, 1999). For this reason, healthcare students practice their interviewing skills through interactions with standardized patients (Barrows, 1993). Standardized patients are actors trained to play the role of a patient. However, due to the expense of hiring and training actors, students get few of these standardized patient interactions (Parsons et al., 2008; Tamblyn et al., 2009). This limitation is one of the reasons educators in the healthcare fields want to provide virtual patients before students interact with standardized patients or real patients.

2.1.2 Virtual Patient Simulation

Educators use virtual patient simulation as preparation for standardized patient and real patient interviews. Just as flight simulators help prepare pilots for real flight, interpersonal simulators, such as virtual patient simulators, help prepare healthcare students for real interpersonal interactions (Johnsen, 2008; Kotranza, 2009; Raij, 2009). Several studies have shown that virtual patients can be used to evaluate and improve cognitive and behavioral skills as well as or better than traditional methods (Huang et al., 2007; Johnsen et al., 2007; Kamin et al., 2002; Kotranza et al., 2009b; Leong et al., 2003). Virtual patient simulation can bridge the gaps in healthcare student education by exposing them to patient disease states that they may otherwise not experience and provide a safe environment to make mistakes (Nutter and Whitcomb, 2001).

Virtual patient simulation facilitates interpersonal skills training using lifelike clinical scenarios in which the user becomes the healthcare professional and practices procedural, diagnosis, and communication skills. Virtual patients present a condition (stomach ulcer, breast cancer, depression, etc.), and the goal of the user (healthcare

student) is to 1) diagnose the patient's condition; 2) prescribe a treatment plan; and 3) develop rapport with the patient through empathy, professionalism, and proper procedure.

The term "virtual patient" covers a variety of computer-based simulations designed to complement clinical training. The majority of existing virtual patient systems are web-based text and video systems (Hayes and Lehmann, 1996; Huang et al., 2007; Leong et al., 2003; Shah et al., 2008). Healthcare students interact with virtual patients by: typing, choosing from a predefined list, or speaking utterances to the virtual patient (e.g. "How long have you had the pain?") (Benedict, 2010; Bergin and Fors, 2003; Ellaway and McGee, 2008; Hubal et al., 2000; Johnsen et al., 2007; Kenny et al., 2008). The virtual patient is represented either as recorded videos of an actor, animated videos, still images, or a 2D or 3D animated virtual character. Virtual patients are displayed on a monitor, in a web-browser, or life-size with a projector, large screen TV, or head-mounted display. In this dissertation, we focus on improving typed and spoken natural language conversations with virtual patients, regardless of representation or presentation.

Evaluation of virtual patient applications has shown educational benefit, but they are costly to develop, which makes them available to few medical schools (Huang et al., 2007). The successful projects have been developed with significant collaborative effort from both domain experts and computer science experts (VH developers). The creators of these VHs report that it is logistically difficult and time consuming to create the necessary conversational models (Dickerson et al., 2005; Kenny et al., 2007; Kenny et al., 2008; Triola et al., 2007; Villaume et al., 2006). A survey of 142 US medical schools

reported that 80% of virtual cases cost more than $10,000 and have a median production time of 17 months; further, these virtual patient cases tend to have limited racial and ethnic diversity, and few cases are produced (Huang et al., 2007). The proposed methods will make the creation of virtual human patients faster and less costly.

2.1.3 The Interpersonal Simulator

The conversational models created using the methods described in this dissertation are compatible with many output mediums (see section 3.2.7) and were specifically designed to facilitate VH interactions using the Interpersonal Simulator (Johnsen et al., 2006; Rossen et al., 2009). The Interpersonal Simulator allows users to interact naturally (speech and gestures) with VHs (Johnsen, 2008). The Interpersonal Simulator has been developed as a collaborative effort of many researchers, and is not claimed as a contribution of this dissertation (Deladisma et al., 2007a; Dickerson et al., 2005; Johnsen, 2008; Kotranza, 2009; Lind and Lok, 2006; Peden et al., 2011; Raij, 2009; Stevens et al., 2006). The work described in this dissertation provides enhancements to the applicability of the Interpersonal Simulator to real-world interpersonal skills education.

An overview of the Interpersonal Simulator setup is shown in Figure 2-1. Speech interaction is enabled by a spoken language dialog system. The system uses a custom speech recognizer based on the Microsoft Speech API called *Exact and Dictation Speech Recognizer* for continuous speech recognition. Realistic VH body meshes have been provided using a variety of software including Autodesk's Maya, Di-o-matic's Facial Studio, and Autodesk's Evolver. The VHs are rendered and animated using the open source Object-oriented Graphics Rendering Engine (OGRE). The characters

employ both skeletal animation (arm movements, leg movements, and head turning) and morph animation (breathing, blinking, lip synching, and emotional expressions). They also employ a realistic gaze model derived from research on controlling user impressions with gaze (Fukayama et al., 2002). The VHs are displayed life-sized using a projector or large screen display.

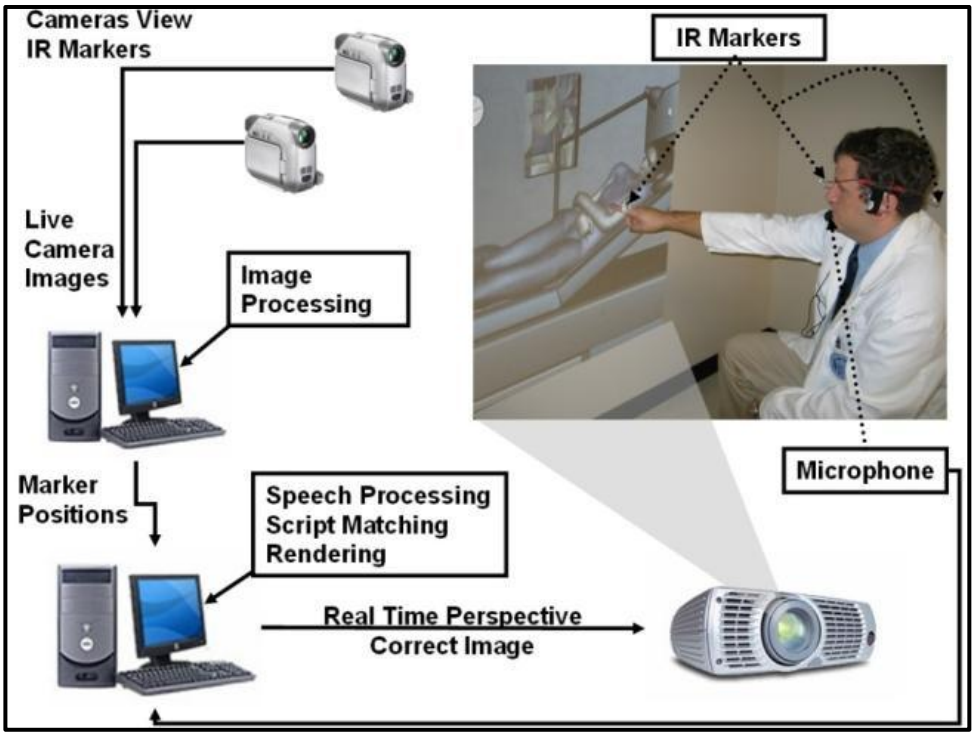

Figure 2-1. Overview of the Interpersonal Simulator (Johnsen, 2008).

The Interpersonal Simulator was validated for use in medical interpersonal skills education (Deladisma et al., 2007b; Dickerson et al., 2005; Johnsen, 2008; Johnsen et al., 2007; Raij et al., 2007). By combining the Interpersonal Simulator with a sensor enhanced mannequin, researchers found the VH interactions could also improve psychomotor and affective performance in a clinical breast exam (Kotranza, 2009; Kotranza and Lok, 2008; Kotranza et al., 2009b). Further, researchers have found that interactions with VHs can elicit racially biased behavior that is predictive of real world

racial biases (Rossen et al., 2008); and after-action reviews can facilitate self-reflection on biases displayed during the interactions (Raij, 2009; Raij et al., 2009; Raij and Lok, 2008). Detecting and reflecting on one's biases is a step towards changing biases, and indicates that VH interpersonal skills may be useful for diversity training (Raij, 2009; Rossen et al., 2008).

These positive findings have created a demand for more VH training simulations. For many interpersonal training goals, tens to hundreds of VHs will be required (Plant et al., 2005; Triola et al., 2007). However, using the previous conversational modeling methods, the time and monetary costs required for the creation of VH conversational models makes these training goals infeasible.

2.1.4 Challenges in Modeling Healthcare Training Conversations

The creation of a virtual patient requires significant collaborative effort between healthcare educators and computer scientists (knowledge engineers) (Dickerson et al., 2005; Glass et al., 2005; Kenny et al., 2007; Kenny et al., 2008; Villaume et al., 2006). As computer scientists, we rarely have the knowledge to create healthcare scenarios on our own. Healthcare educators have that knowledge; however, educators do not have the technical expertise necessary to create virtual patients capable of natural language conversations. A system that enables educators to construct virtual patients for healthcare interview training without computer science expertise would enhance development efficiency and expand utilization.

Existing systems which allow creation of virtual patients by healthcare educators are restricted to multiple choice interactions or structured queries (Benedict, 2010; Fall et al., 2005). In multiple-choice interactions, students choose their questions and statements from a predefined list. In structured query interactions the users are not

given the predefined list directly, each word typed into a text box brings up a menu of questions they can ask that contain the word. These predefined list interactions focus on the fact-finding mission in order to reach a diagnosis.

In contrast, the interpersonal simulation systems addressed by this dissertation use natural conversation to train interviewing skills. Improving interview skills requires practicing those skills, and natural interaction performs better than linear or forced branching for training those skills (Bearman et al., 2001; Johnsen et al., 2007; Saleh, 2010; Yedidia and Lipkin, 2003). Using natural conversation, the student interviews the virtual patient by asking questions about present health and past medical history in the student's own words, in any order.

Prior to the work described in this dissertation, the researchers who developed the Interpersonal Simulator created three virtual patients using CCM. These first three virtual patients were created over the course of four years. The creation of each virtual patient using CCM required approximately 200 hours and over 6 months.

Because of the limited number of cases created, these virtual patient systems have been useful in a narrow scope. To provide wide-spread functional benefit to medical schools, tens to hundreds of virtual patient scenarios are necessary, resulting in prohibitive time and cost requirements (Huang et al., 2007; Triola et al., 2007). The proposed conversational modeling methods reduce this time significantly, and thus increase the potential for virtual patient based curricula.

2.2 Conversational Modeling

There are several aspects to the creation of conversational VHs, including:

- Visual: how virtual humans appear

- Non-verbal: face and body animations, tracking the user, eye gaze and other non-verbal modes of communication
- Conversational: utterances that VHs understand and how VHs respond

In this dissertation, we focus on the verbal or conversational aspect of VH creation. VHs conduct natural language conversations (as opposed to multiple choice) using un-annotated corpus retrieval approaches and are primarily question-answering agents (Dickerson et al., 2005; Kenny et al., 2007; Kenny et al., 2008; Leuski et al., 2006; Leuski and Traum, 2010; Roque and Traum, 2007). The corpus of a VH consists of stimulus-response pairs of what the users will say to the VH (stimuli/questions) and what the VH will say back (response). When the user asks a question, the system searches the corpus for the most similar question and provides the paired answer using keyword vector matching or statistical distribution methods (Dickerson et al., 2005; Leuski and Traum, 2010). For example, if a user asks the virtual patient “How are you feeling?”, “What’s wrong?”, or “How can I help you?”, the virtual patient will respond with “I’m having stomach pain”.

A question-answering VH is capable of simulating a healthcare patient for a medical interview (Johnsen, 2008; Raij, 2009). This is possible because of the structure of the medical interview itself. The interview occurs within a restricted domain and is driven by the interviewer (medical student). The restricted domain allows a corpus to enumerate the space of a conversation. However, if the interviewer discusses a topic outside of the restricted domain, it is likely that the VH will not have a response, or will respond incorrectly. In the interview, the interviewer asks questions, and the VH patient responds. This is a simplified interaction when compared to human-human interviews. While human-human interviews are still primarily question-answering, the human patient

may provide backchannel information that is limited or non-existent in a purely question-answering conversation. Despite this simplification, the VH patient simulation is able to provide beneficial question-answering interactions.

The methods described in this dissertation are used to generate VHs that conduct conversation-specific question-answering or question-asking interactions (Rossen et al., 2010; Rossen et al., 2009). We define conversation-specific VHs to be those designed to know only what they need to know for their particular conversation. They do not search for additional information or ad lib, they remain in character to provide a realistic interpersonal experience (Dickerson et al., 2005; Leuski et al., 2006; Rizzo et al., 2010).

2.2.1 The Conversation-specific Problem Space

Conversation-specific agents are distinct from open-domain and domain-specific conversational agents. Conversation-specific VHs focus on providing an interpersonal experience of a specific conversation, with the goal of facilitating experiential learning (Itin, 1999; Johnsen et al., 2005). The methods proposed in this dissertation address the problem of acquiring knowledge for conversation-specific conversational models.

Open-domain conversational agents focus on passing the Turing Test (appearing indistinguishable from a human conversational partner) and not on delivering information (Mauldin, 1994; Wallace, 2005). Their goal is believability as a conversational partner and they often employ conversational tricks to accomplish this goal; such as repeating statements back as questions or using distracting statements (Colby et al., 1971; Weizenbaum, 1966). Domain-specific conversational agents typically focus on helping users complete a task such as information search, travel booking, or diagnosis (Goh et al., 2007).

We propose a definition of the conversational modeling problem space illustrated in Figure 2-2. Conversational modeling is a subfield under the field of Artificial Intelligence. Within that subfield, there are modeling approaches that depend on the level of domain specificity. Each level of domain specificity has a certain goal. The conversational modeling methods described in this dissertation address one type of domain specificity, conversation-specific, and address one of the problems with modeling conversations at that level of specificity, knowledge acquisition.

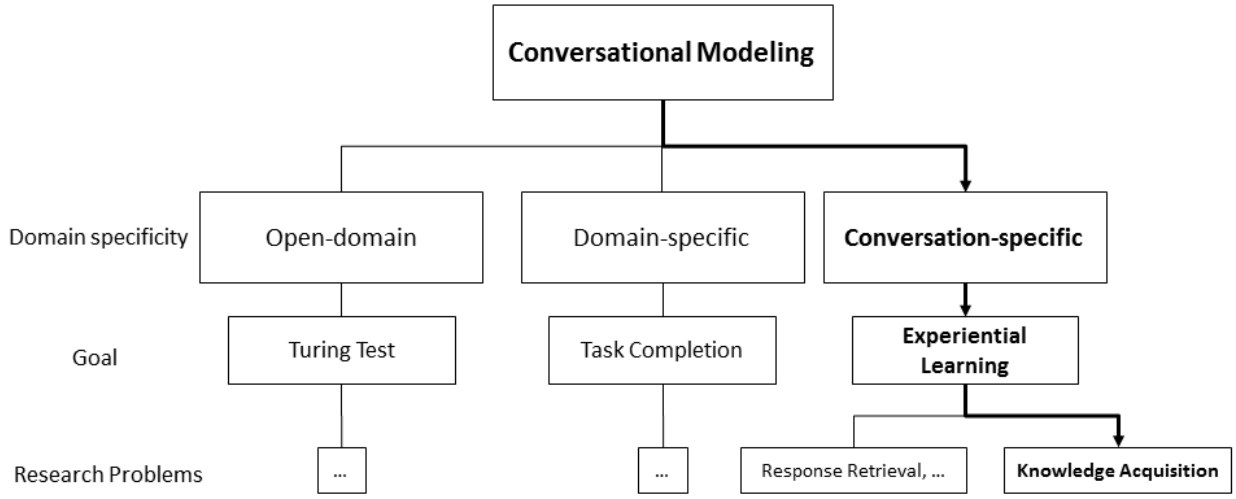

Figure 2-2. The problem space definition. Bold lines lead to the problem space addressed by the conversational modeling methods in this dissertation.

2.2.2 The Challenges of Knowledge Acquisition

The creation of conversation-specific VHs versus open-domain and domain-specific conversational agents differs in that conversational tricks are counter-productive and the data for specific conversations does not exist on the internet or in other easily accessible corpora resources. Conversation-specific VH authors have to collect data themselves (Ruttkay et al., 2004). The majority of incorrect responses in this type of conversational model come from unanticipated topics of conversation (Dickerson et al.,

2005). This symptom of unanticipated topics indicates a problem with knowledge acquisition for creating conversation-specific conversational models.

In order for conversational models to respond accurately to a wide variety of inputs, they require a corpus that enumerates the stimulus-response space of a conversation (Dickerson et al., 2005; Reiter et al., 2003; Rossen et al., 2009). Using CCM, VH authors have encountered the following challenges to creating corpuses that enumerate the stimulus-response space of a conversation:

Challenge 1: Experts lack knowledge detailed enough for generalization. Accurate conversational models require thousands of stimuli paired with hundreds of responses (Chapter 3, 4). Knowledge based on recordings of “natural” or staged interactions and asking experts is not detailed enough for generalization (Reiter et al., 2003). Experts are likely to come up with a small fraction of the required number of stimuli. For example, in the study described in Chapter 3, the pharmacy educator was unable to anticipate the 174 syntactical ways to ask the 53 semantically unique questions about Aspirin. The educator was able to predict only 10 of these semantically unique questions. In a previous study, we have seen that unanticipated stimuli account for the majority of errors (51%) in a conversation modeled using CCM (Dickerson et al., 2005).

Challenge 2: Experts phrase questions and conduct interviews differently than novices do. The target user group for interpersonal skills training is domain novices. Domain experts do not phrase questions the same way novices do. With more experience, experts use shorter, more focused questions and infer information from their past experiences (Westberg and Jason, 2001). Experts also conduct shorter, more

focused interviews, and so cover few of the topics that a novice interviewer will cover. The effect of these differences is that if VH authors acquire knowledge from experts, the corpuses do not cover the questions asked by novices.

Challenge 3: Limited use of existing corpora sources and costly creation of new corpora resources. Because of the above challenges of acquiring knowledge from expert sources, in order to enumerate the space of a conversation between a VH and a novice, VH authors must acquire knowledge from material originating from novice sources. These novice sources have logistical issues regarding legal use of existing material, monetary cost, required time, and end-user availability. For many human-human interactions, there are legal restrictions regarding viewing, particularly in healthcare; the problem being that non-healthcare professionals are not allowed to view real patient interviews. Even with staged interactions, such as students interviewing actors, there are the logistical difficulties of hiring and training actors, issues in standardization and repeatability, as well as monetary costs. Wizard of Oz (human-controlled) VH interactions also have the same drawbacks as using actors in terms of availability, compensation, and standardization.

Challenge 4: Extracting knowledge from corpora sources. An additional problem with novice sources is that of extracting utterances from interactions. After each set of user interactions, the knowledge engineers review the videos and transcripts from those interactions to extract new stimuli and correct gaps in the corpus indicated by the transcripts. Knowledge engineers determine if non-responses were caused by missing stimuli or speech recognition errors; and if incorrect responses were caused by erroneous similarity between existing stimuli, missing stimuli, or speech

recognition errors. Acquiring knowledge from standard novice sources has a time cost for both knowledge engineers and collaborators.

Challenge 5: Limited expertise availability and costly collaboration. The fifth challenge is one of collaboration. Knowledge engineers may not know the domain, so they must collaborate with domain experts to validate the stimuli and create new responses (Sutton et al., 1996). Before knowledge engineers can begin working on a domain specific VH, they need to learn about the domain from an expert. Even after this education, they are not experts themselves. This means that the knowledge engineers will need to contact the domain expert (e.g. a medical doctor, psychologist, military expert, etc.) every time they want to a) validate a new stimulus; or b) create a new response to a stimulus. This collaboration takes time, and there are often communication challenges because of differing backgrounds.

In practice, these five challenges result in inadequate initial knowledge in the conversational model, few iterations of user testing, and iterations of testing having a limited number of users. Thus, the resulting conversation corpus has significant gaps in its stimuli coverage. This causes increased response errors and a decreased ability for the VH interaction to achieve educational and training objectives.

The methods proposed in this dissertation, HDCM and Conversational Knowledge Reuse, address these challenges. HDCM and Conversational Knowledge Reuse are new approaches to rapidly acquiring knowledge directly from novices and experts. They remove the manual identification of utterances from video by knowledge engineers and they create faster collaboration through a distributed system.

2.2.3 Addressing the Challenges of Creating Conversational Corpora

The cost of developing conversational models is large in part because of the required size of a corpus to facilitate robust conversations. These conversations require a considerable amount of knowledge about potential speech inputs and outputs. Generalization requires a large data set (corpus) that covers the unusual boundaries of realistic inputs; that is, inputs that real users would say to a VH.

While expert knowledge cannot provide the corpus with realistic novice inputs, expert knowledge *is* a good way to validate inputs for this corpus and to produce outputs (Reiter et al., 2003; Rossen et al., 2009). Experts can identify if questions are on topic, and can create patient responses to those questions. HDCM accepts these constraints by using a collaborative authoring tool that collects knowledge from novice users, which is validated by the expert users (see Chapter 3 for details).

The idea of engaging end-users for knowledge acquisition was explored in *Open Mind Common Sense* (Singh et al., 2002). The goal of *Open Mind Common Sense* is to build software agents that are capable of common sense. The project uses an online tool for collaborative knowledge acquisition. Their approach is similar to the construction of other collaborative web-based efforts, such as the Open Directory Project or Wikipedia. The contributors for these projects are motivated to improve the project itself. While these projects have found great success, their approach would not work for communication skills training applications -- students are not motivated to engage directly in the process of modeling VH conversations for their own training materials, so HDCM engages them indirectly (Villaume et al., 2006).

We find a solution for engaging our novice users in Lois von Ahn's *ESP Game* (von Ahn and Dabbish, 2004). Von Ahn pointed out that human-based computation can

solve problems that are still untenable for computers to solve, e.g. tagging images for searching. In the *ESP Game*, online players guess what their game partner is looking at by naming parts of an image. They are motivated because the game is fun. Google has used this game to tag huge numbers of images, thus letting Google search images without processor intensive vision techniques. We build upon this work by taking and extending the “human computation” (human-centered) approach to knowledge acquisition for VH conversations. We use interactions with novices to acquire a corpus of realistic input data. The important lesson from the *ESP Game* is to set up the task so that users accomplish their own goals (learning) in a way that causes a beneficial side effect. In the case of conversational interactions, that side effect is teaching a VH to conduct a conversation.

In the following chapters, we present our solutions to the various problems of acquiring knowledge for interpersonal simulation.

CHAPTER 3
HUMAN-CENTERED DISTRIBUTED CONVERSATIONAL MODELING

This chapter presents the analysis, design, implementation, and evaluation of Human-centered Distributed Conversational Modeling (research questions 1-3). It also describes an implementation of Human-centered Distributed Conversational Modeling, Virtual People Factory. It goes on to report a meta-analysis on evaluations of Virtual People Factory in four real-world applications. The initial evaluation of this method and implementation was published in the proceedings of the International Conference on Intelligent Virtual Agents (Rossen et al., 2009). An extended version discussing additional details on the method and implementation, as well as a meta-analysis of publications by healthcare professionals using Human-centered Distributed Conversational Modeling and Virtual People Factory, was accepted to the International Journal of Human Computer Studies and is awaiting publication (five year impact factor = 2.3) (Rossen and Lok, 2012).

Personal Contributions. I conceptualized the Human-centered Distributed Conversational Modeling method, implemented the Virtual People Factory system, and designed and analyzed the initial study. I also collaborated with healthcare practitioners using Virtual People Factory on several of their studies and publications.

Collaborators. Dr. Kyle Johnsen, Dr. Andrew Raij, Dr. Aaron Kotranza, and Scott Lind, M.D. were involved in the discussions of the Human-centered Distributed Conversational Modeling method and provided valuable reviews and feedback during the development of Virtual People Factory. For Evaluation 1 (below), Dr. Carole Kimberlin provided access to study participants (pharmacy students) and served as a participant herself by re-creating the Vic Johnson Dyspepsia virtual human patient.

Many other healthcare professionals were involved in the additional studies using Human-centered Distributed Conversational Modeling and Virtual People Factory, and their publications are referenced in Evaluation 2 (below).

Relevance to thesis. This chapter reports several studies indicating that educators are not only using Virtual People Factory in healthcare education, but they also feel it is providing educational benefit. It provides evidence that Human-centered Distributed Conversational Modeling reduces the time to model virtual human conversations, and because of that reduction we have seen an increase in both the number of virtual humans created and in the diversity of the applications to which those virtual humans are being applied.

3.1 Overview

Human-centered Distributed Conversational Modeling (HDCM) applies the ideas of crowdsourcing and human-based computation to the challenges of conversational modeling in order to alleviate the bottlenecks of Centralized Conversational Modeling (CCM). We see in section 1.1 Problem Statement and section 2.2.2 The Challenges of Knowledge Acquisition, that the knowledge engineer's role in creating the conversational model is collecting knowledge from the experts and novices, and using that knowledge to generate a machine readable corpus. We can remove these duties from the knowledge engineer by providing a guided learning system for use by the experts and novices to "teach" the system directly. This process is referred to as human-centered because it fits the way domain experts think about creating a virtual human (VH), and is a natural method for domain novices to participate in the VH creation process.

The process of HDCM is collaboration between domain novices (the learners) and a domain expert (the educator) to teach a VH how to conduct a conversation. The expert enters an initial set of questions and responses. This set is the outline of the conversation and seeds the VH's learning process. The goal is to minimize the expert's upfront cost of creating a VH by allowing the conversation to grow through iterative refinement.

Next, the expert enlists the help of novices. The novices attempt to conduct a conversation with the VH. The VH will perform poorly during this first interaction by either not having a response to a question, or by responding incorrectly. These errors are logged and are later displayed for the expert one at a time.

The expert then enters new responses to each new stimulus, or matches new stimuli to existing responses. After all new stimuli have been processed and all the new responses have been added to the conversational model, the expert initiates a second iteration. They send the interaction to a larger group of novices. After a few iterations of this process, the expert will start to receive diminishing returns -- each new interaction produces fewer and fewer new stimuli (see section 3.3.1 for details and 4.4.4.3 for additional evidence of diminishing returns) (Rossen et al., 2012; Rossen and Lok, 2012).

The end-condition for this process is dependent on the complexity of the conversation and the required accuracy. For a VH that needs to discuss few topics and those topics are straightforward in nature (e.g. "What is your name?", "What is your age?" etc.) the process requires fewer participants than for complex topics such as a family history of cancer, where stimuli content overlaps ("How did your father die?",

“How did your mother die?”, “How old was your father when he died?”). Complex topics require more distinguishing factors and more iterations of testing will be necessary with greater numbers of users in each iteration. As a data point, in the evaluation of the relatively complex scenario described in section 3.3.1.1, for a 20 minute conversation to achieve 79% accuracy required three iterations consisting of a total of 186 participants.

HDCM’s guided learning system uses an approach known as case-based reasoning. Case-based reasoning’s defining element is the reuse of information from specific previous experiences to come up with a *response* for the current *stimulus* (Aamodt and Plaza, 1994). Case-based reasoning systems learn by identifying successes and failures in order to solve similar problems in the future. In the context of conversational modeling, the stimuli are user questions/statements, and responses are VH speech. Failures consist of either the VH lacking a relevant response, or the VH response being incorrect. Once a failure is identified, the expert enters a correct response so that the system can achieve success in the future.

Using HDCM, domain experts and novices asynchronously collaborate to teach the VH how to converse. They collaborate through a graphical user interface that is useable without any knowledge of the technical details of conversational modeling, such as XML or case-based reasoning. Figure 3-1 shows the iterative process end-users follow for creating a VH conversational model and is described in more detail below.

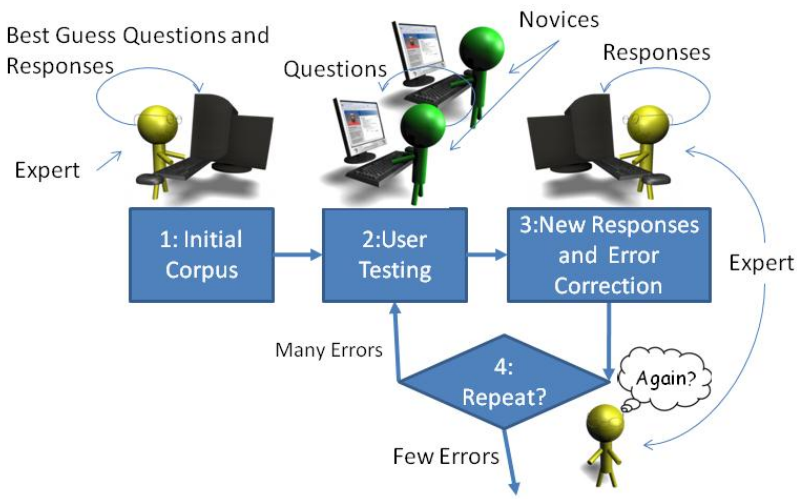

Figure 3-1. The Human-centered Distributed Conversational Modeling process.

Phase 1: A domain expert seeds the VH Conversational model with their best guesses as to what will be said to the VH and what the VH should say back.

Phase 2: Multiple novices have a typed conversation with the VH. The system collects new stimuli when the VH does not have a response, and when it responds incorrectly.

Phase 3: A domain expert enters responses to which the VH could not respond, or to which the VH responded incorrectly.

Phase 4: Phase 2 and 3 are repeated until an acceptable accuracy is reached. In practice, the acceptability of the accuracy is determined by the domain expert.

Through interactions with a VH, the domain novices enumerate the space of what will be said to the VH; while domain experts enumerate the space of what the VH will say back. During interactions with the novice, the system gathers three types of errors – true-negative, false-negative, and false-positive. A true-negative error occurs when a user provides a stimulus, and the system cannot find any response because there is no appropriate response in the corpus. With a false-negative, there is an appropriate

response, but the system fails to match the stimulus to that response. A false-positive occurs when an inappropriate response is given, based on a mismatch of stimulus to an item in the corpus. The system enters these errors into the list of new stimuli. After gathering errors, the expert adds new stimuli and responses to the conversation corpus to facilitate future accurate responses. Compared to CCM, iterations of HDCM are completed faster, and can involve a greater number of end-users.

A major barrier to using VHs in curricula is time (Huang et al., 2007). By shortening the iteration cycle and easing the distribution, we shorten the time to create a VH and increase the potential for these systems to be used in the real world. This process is a learning system that generates a corpus enumerating the space of a conversation. That corpus forms the basis of a VH conversational model for corpus retrieval conversations. The HDCM method is implemented in the web-based application Virtual People Factory (VPF).

3.2 Implementation: Virtual People Factory

VPF is a platform for the development and deployment of VHs using the HDCM method. VPF provides interfaces for both expert and novice users as well as a web-service for communicating with external applications. The application consists of three subsystems:

1 a browser-based interaction system,
2 a VH editor system, and
3 a web-service developer API.

These three parts are used asynchronously to create conversational models and interact with VHs. VPF's conversational models are referred to as *scripts*. Scripts are used in the simulation of a conversation. Scripts consist of both the set of stimuli and responses (corpus) as well as supporting tags such as associated animations,

emotions, audio speeches, and images. VPF supports script creation through the collaboration of novice users and expert users. The novice users have a client interface to interact with the VHs, and the expert users have a set of interfaces for creating and modifying scripts. These interfaces and VPF's learning management system facilitate the collaborative design and deployment of VH scripts (Figure 3-2).

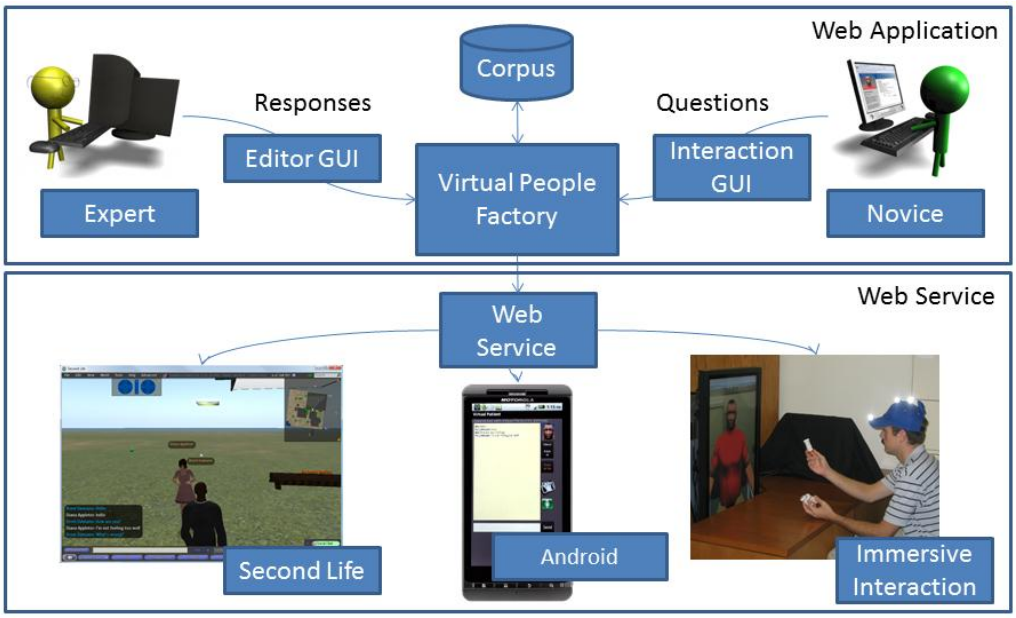

Figure 3-2. Virtual People Factory System Overview – the system is divided into the web application and the web service subsystems. The web application provides browser based clients. The web service supports other client interfaces such as Second Life, Android, and the Interpersonal Simulator.

Definitions:

- Script: a representation of the knowledge of a VH (including the stimuli-response corpus, animation tags, audio tags, emotion tags, etc.)
- Text response: the text of a speech-based response
- Audio file: a tag indicating an audio file to play as part of a response
- Animation: a tag indicating an animation to play as part of a response, as well as the timing of the animation

- Emotion: a tag of an emotional facial expression to make as part of the response, as well as the timing for the movements involved expressing the emotion
- Discovery: a tag indicating an important piece of information contained in the response
- Topic: a tag indicating the subject or theme of the response

3.2.1 Virtual People Factory Server

As shown in Figure 3-3, VPF uses a client-server architecture. The server portion of VPF runs on a single PC server containing a Core2 Duo Quad-Core processor and 4GB of RAM running the Apache Web Server on Microsoft Windows. The VPF Server runs on the open-source software components: PHP Scripting Language, MySQL Database, and the MySQL Ajax Database Access Layer (Rossen, 2010). All communication is performed over http using Ajax calls. Data is marshaled and passed from one application to another in JSON or XML format across http by the *VPF Web-Service API*.

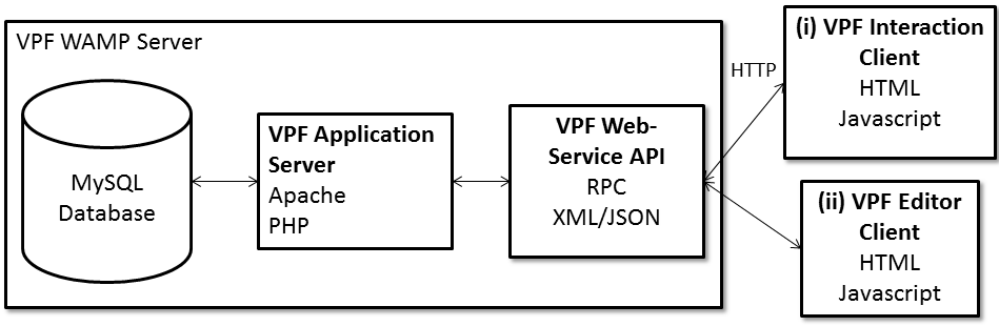

Figure 3-3. The Virtual People Factory Architecture using Windows, Apache, MySQL, and PHP

3.2.2 VPF Web-based Clients

The VPF web clients run on the user's local machine in a web-browser such as Firefox, Chrome, Safari, or IE7+. The client side of VPF is divided into two systems; the

interaction system (i) and the editor system (ii) (Figure 3-3). The VPF interaction and editor interfaces were implemented using html, CSS, and JavaScript with jQuery.

3.2.3 Interaction Interface

VPF's browser-based interaction interface is an instant messaging style interaction with a VH. This design choice was made so that the development of the VH corpus could be conducted independently from development of the visual elements of the VH and the rest of the script. This decoupling promotes starting generation of the corpus early during the development of a new VH. A script can start out with a corpus of just a few questions and responses. Since that is all VPF needs for an interaction, people can immediately start interacting with the character, and developing the corpus using HDCM.

A further advantage of this decoupling is the increased user concentration on the conversational aspects of the interaction. The advantage for developing a corpus is that users try to gather the necessary information from the text if visual information is lacking. Once visual information is added, users expect to be able to gather that information (such as emotions) from the visual features of VHs (Cassell, 2001).

While VPF's interaction can be as simplistic as a text-based instant message interface, it also expands to incorporate additional features such as animations, audio, topics, and discoveries. The animations are represented as a tag in the script rather than an embedded binary of the information. This tagging allows any interface that uses the script to interpret the tag in its own way. For example, the animation "wave" in VPF would be added to the text-based response as (*waves right hand*) while in a 3D interaction the system would play a "wave" animation showing the character raising his hand and waving.

In Figure 3-4 we show the VPF interaction screen. On the left we see the patient's information, including the patient's name, and a description of their case. On the right we see a transcript of an interaction.

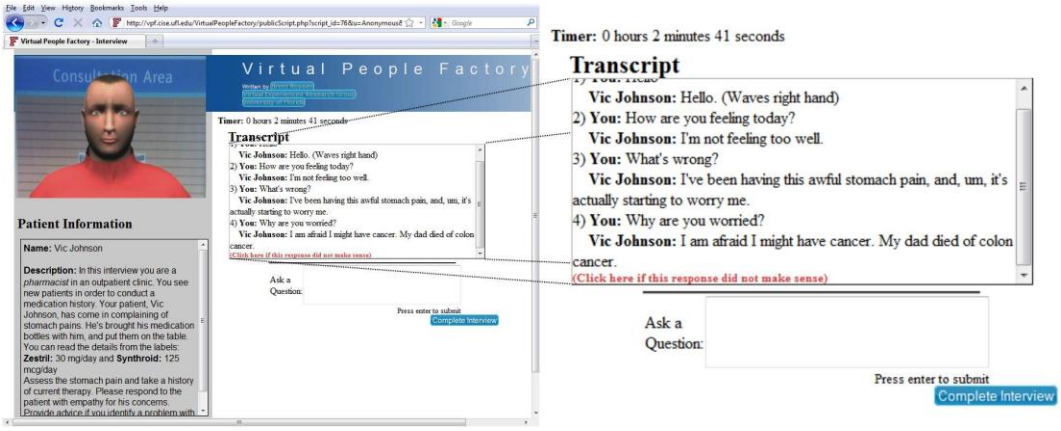

Figure 3-4. The Virtual People Factory browser-based interaction interface

3.2.4 Implementation of Error Gathering

VPF implements the learning process described by HDCM. VPF's responsibilities in this process are gathering errors and facilitating error correction. As described at the end of section 3.1, during interactions VPF gathers three types of errors – true-negative, false-negative, and false-positive. True-negative and false-negative errors, where the VH does not respond at all, are automatically added into a log of new stimuli by VPF. However, VPF cannot reliably identify false-positives. False-positives result from a mismatched stimulus, where the VH did respond, but incorrectly. For example, if the user asks, "Do you take *Tums* regularly?" and the character responds, "I take *Aspirin* all the time." VPF cannot reliably identify that the character's response was about the wrong medication. Accordingly, when the VH responds incorrectly, the instructions ask users to press the "Click here if this response did not make sense" button (seen in red at the bottom of the transcript in Figure 3-4). Pressing that button logs the false-positive

error as a new stimulus for the expert to validate later. After gathering errors, the expert uses the *VPF Editor System* to correct errors by processing the resulting list and adding new stimuli and responses to the conversation corpus.

3.2.5 Editor System

Domain-experts use the *VPF Editor System* to create and edit *VH Scripts*. The *VPF Editor System* has facilities for editing scripts, sharing scripts with other authors, sending scripts to students as VH interactions, and then analyzing and processing the interactions. A design goal of the editing system is to minimize the cognitive load on the user in order to improve the user's ability to learn the system quickly and successfully accomplish tasks (Lidwell et al., 2003). To this end, users start-off with access to the basic features, but can request access to more advanced features -- providing access to only the basic features at first helps users to get started quickly with creating a question-response VH. The advanced features allow users to divide the script into acts, add audio, add free-form xml, and add animation tags. These advanced features provide facilities for use in non-VPF interactions such as Second Life and the Interpersonal Simulator (see section 3.2.7 for additional details).

VPF has multiple methods to input questions and answers. The manual way is using the *Edit Scripts* interface. On this page, users see the list of speech responses, and the set of questions (stimuli) that will trigger those responses. Experts annotate these responses with animations, audio, emotions, discoveries, and topics.

The recommended way for authors to seed new question-response sets is to converse with the character using the *Test Script* interface. The *Test Script* interface allows the Script Editor to conduct a conversation with their VH within the *VPF Editor System*. During the *Test Script* conversation, when the VH does not have a response,

the expert can immediately enter a new response, or connect the stimulus to an existing response. Experts play both sides of the conversation and seed the conversational model for future interactions with students.

After the conversational model is seeded, the expert sends out invitations for students to interview the VH. These invitations are automatically generated using VPF's *Groups System*. The *Groups System* allows experts to add students to the system, and then track their progress. As the students perform interactions, the educator-expert can view transcripts and analyze performance. These student interactions also improve the conversational model. Each time a new utterance is encountered; this information is stored in the *Suggestions System*. The overall flow of knowledge is shown in Figure 3-5.

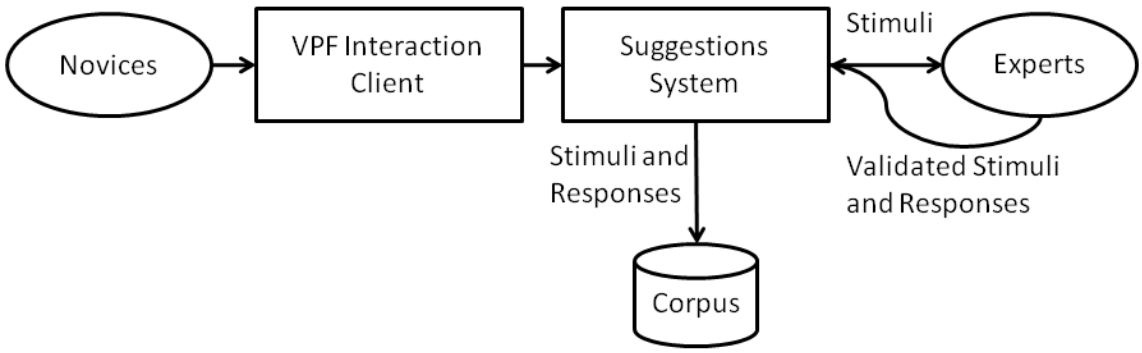

Figure 3-5. The flow of knowledge in the Virtual People Factory system implementation of Human-centered Distributed Conversational Modeling

3.2.6 Suggestions System

The *Suggestions System* displays new stimuli to the expert one at a time. Since these conversational models often grow to include hundreds of responses, it becomes difficult to recall the correct response for a given stimulus. To alleviate this problem, the suggestions system provides help in selecting appropriate responses that already exist in the conversational model.

A screenshot of the interface is shown in Figure 3-6. In this example, for the new stimulus, "Have you been distracted?" the system has provided a list of likely responses (the rightmost list in the image). The user has selected one of these likely responses ("I am easily distracted"), and the system has provided a list of similar responses. The list of similar responses is used when the script editor would like to select a different, but similar, response. The user can also use free-text to enter a new response, and the system will find similar responses.

Once the user presses the "Make Change" button, the system connects the new stimulus to the existing response, or if a new response was entered, the new stimulus-response pair is added to the script. This design leverages the greater ease and speed of recognition memory over recall memory to improve the experts efficiency in processing new stimuli and improving the conversational corpus (Lidwell et al., 2003). Using this interface, script editors can rapidly process new stimuli and either connect them to existing responses, or add new stimulus-response pairs into the conversational model (details on the efficiency of this interface shown in section 4.4.4.3).

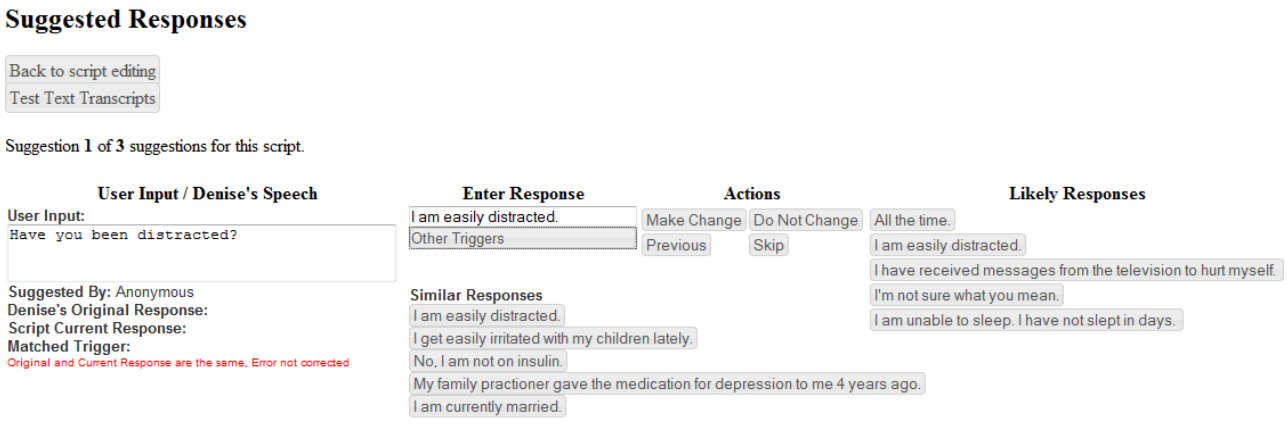

Figure 3-6. The Suggestion System interface – "User Input" is the utterance spoken by the interviewer, "Enter Response" is the space to enter a new response for the virtual human, "Similar Responses" are populated as the author types in a response and fills in the response when selected, and "Likely Responses" are

a list of the relevant responses in the corpus. The author adds the new stimulus and response pair to the corpus by pressing “Make Change”, or ignores the suggestion by pressing “Do Not Change”.

3.2.7 VPF Web-Service API

VPF is intended as a backend system for generating VHs, and then deploying those VHs using a variety of clients. For this reason, VPF was written using client-server principles, and provides a web-service API for interfacing with the conversational engine. Through the *VPF Web-Service API*, VPF VHs can be deployed using a variety of clients. With this technique, users can experience the same VH conversation using different interfaces.

The *VPF Web-Service API* was implemented by extending a MySQL Ajax Database Access Layer web-service (Rossen, 2010). In a standard MySQL Ajax Database Access Layer web-service, each table of the database is represented as its own class. These classes use object relational mapping to allow clients of the web service to access any VH data in the database (after security authentication). The basic web-service was extended to provide a conversational simulation API, including a speech understanding service and transcript service. Through these services, the *VPF Web-Service API* allows VPF VHs to be used as the backend simulation system for a variety of VH interfaces. Each interface uses the same conversational engine. So far, VPF has been integrated with four interface clients – the web browser interface (described above), Android mobile platform, Second Life, and The Interpersonal Simulator.

The VPF Android interface simulates VH interactions using text and voice inputs. Users speak into the phone, and receive both text and audio responses. Currently, the characters’ visuals are still-image only. A small screenshot of the interface is shown in

Figure 3-2, above. The VPF Android interface can be downloaded from the Android Marketplace under the name *Virtual Patient*.

VPF also supports distribution of VHs using Second Life through the *VPF Second Life* application. Using VPF with Second Life provides a body for the VH, support of VH interaction with many users, provides users with their own avatar, and supports user created 3D content (Ullrich et al., 2008). The use of VPF Second Life requires no programming or XML editing. *VPF Second Life* is a middleware application; VPF VHs are loaded into Second Life by providing a VPF username, password, and script information; Second Life username, password, and a location in Second Life for the VH to appear; and then pressing “Log In”. *VPF Second Life* then uses libopenmetaverse in combination with the VPF Web-Service API to connect Second Life with the VPF server (Freedman et al., 2010). The *VPF Second Life* application can be downloaded from the VPF home page after login. The *VPF Second Life* application is used by healthcare practitioners at the University of South Florida to provide virtual clinical skills labs, where VPF characters are constantly available for healthcare student practice (Jackson, 2010).

The Interpersonal Simulator allows users to interact with a life-size VH and conduct natural language spoken conversations (section 2.1.3). Users walk up to these characters and speak; when they speak, an automatic speech recognizer transcribes their words to text. The text is translated into XML queries and sent to the *VPF Web-Service API*. VPF provides speech and animation responses, which the Interpersonal Simulator renders to an immersive environment.

In addition to the current set of clients, VPF's flexible backend can support future VH interfaces as well. These future interfaces will be created using the same VPF backend through the web-service API. Using the same backend for many front-ends promotes reuse of conversational models created using one client interface to be leveraged for many client interfaces.

3.3 Evaluations of Human-centered Distributed Conversational Modeling and Virtual People Factory for Healthcare Interview Training

We evaluated HDCM to establish the efficacy of this method for generating conversational models. We further examined if VPF can be used for HDCM by experts in real-world practice without assistance from the VPF developers. Last, we established the usability of VPF and examined limitations of the current version of VPF. The goal of these evaluations was to understand the impact of the HDCM method on conversational modeling and establish if the current implementation of VPF is usable in real-world educational settings.

We separated the evaluation into three parts:

- Evaluation 1: a case study evaluation of HDCM and VPF on the creation of one conversational corpus (section 3.3.1)
- Evaluation 2: a meta-evaluation of 4 published case studies that further examine student and educator real-world experiences with HDCM and VPF (section 3.3.2)
- Evaluation 3: a usability evaluation based on self-reported feedback from experts on the efficacy of HDCM and VPF in real-world educational settings (section 3.3.3)

We evaluate HDCM using an in depth examination of the creation of one conversational corpus. We further report on the creation of four additional conversational corpuses and their efficacy for use in healthcare education. Last, we discuss self-reported expert feedback on the usability of VPF to create virtual patients and educate students.

3.3.1 Evaluation 1: Speed of creating a virtual patient

In this evaluation, we examine if the HDCM approach enables experts to create conversational models, reduces conversational modeling time requirements compared to CCM, and results in a conversational model with increased accuracy for spoken interactions.

3.3.1.1 Methods

To evaluate HDCM, a Dyspepsia (discomfort centered in the upper abdomen) conversational model was developed for an Introduction to Pharmacy Communications course taught in spring of 2008 at the University of Florida College of Pharmacy. The character for this scenario is named Vic. At minimum, Vic needed to discuss the following topics: Chief Complaint of stomach pain, Age, Weight, Gender, Blood Pressure Readings, Thyroid Readings, Fears of Cancer, Risk Factors (Smoking, Alcohol, Drugs, Allergies), Medical Problems (Hypertension, Hyperthyroidism, Back Spasms), Medications (Zestril, Synthroid, Aspirin, Tums), and his parents' Medical History (Father died of colon cancer, Mother died of a heart attack). Vic needed extensive domain specific knowledge in order to converse about these topics. This scenario was previously generated using CCM, and the original conversational model is used for comparison (Johnsen, 2008).

In the current study, the pharmacy instructor (domain expert) and pharmacy students provided domain knowledge using the HDCM process. Here is the HDCM process the pharmacy expert followed to create Vic (generic version shown in Figure 3-1, above):

Phase 1: The Pharmacy Instructor used Vic's required topics to create the initial set of questions and responses. To do this, she played the role of a student and asked Vic questions, and responded to those questions herself.

Phase 2: Pharmacy teaching assistants and students interacted with Vic through VPF's web-based interface. This identified missing stimuli for which Vic did not respond, or responded incorrectly.

Phase 3: The Pharmacy Instructor added new responses for the new stimuli, or connected the new stimuli to existing responses.

Phase 2 and 3 were repeated three times.

Participants. Two classes of participants were involved in the study, a domain expert, and two types of domain novices:

Domain Expert: the pharmacy instructor had standard computer experience with word processing and email. The instructor was motivated to participate by a desire to give her students "early practice experiences".

Domain Novices: the pharmacy instructor recruited pharmacy students from her Introduction to Pharmacy Communication Skills course. The participants consisted of 12 teaching assistants (TAs) and 174 second-year pharmacy students. Participant ages ranged from 20 to 60 with an average of 25.44. The pharmacy students received extra credit in the course for interviewing Vic for a minimum of 10 minutes and 25 questions.

Procedure. The pharmacy instructor uploaded a comma-separated list of student names and emails into the VPF Groups system; VPF generated customized links for each student, and sent out emails. Students went to the website where they completed a consent form, conducted a typed interview, and completed a post-interview questionnaire. Participants had two weeks to conduct their interview at their own convenience.

Data Analysis. Data analysis was divided into three parts: 1) conversational modeling time; 2) conversation accuracy improvements; and 3) accuracy in comparison to a previous CCM model. Part 1 established the amount of time required to model a

conversation using HDCM and VPF. Part 2 established the trend of changing accuracy during each iteration of HDCM. And part 3 established if the resulting HDCM conversational model is more accurate than the previously created CCM conversational model.

The HDCM conversational model was created using interactions with the TAs and students of the Introduction to Pharmacy Communications Skills course. Users were divided into three iterations of model improvement, the first 12 teaching assistant participants (group TA), the next 44 student participants (group S1), and the remaining 130 student participants (group S2). The 12 TA interactions were conducted prior to student interactions in order to seed the system and provide a more developed system for the first round of students. The two student groups were divided based on when the expert processed the first set of student suggestions. The expert processed the first set of suggestions near the end of the first week of the study, and continued to process suggestions throughout week two. Since students participated online, and at their own convenience, the grouping of participants into S1 and S2 was self-selected. The students who chose to participate in the first week were included in S1; the students who chose to participate in the second week were included in S2. In section 3.3.1.2, we show the number of unique questions and responses acquired from these interactions, as well as the percentage of accurate responses with each group of users. We further evaluated this conversational model for accuracy with spoken inputs using transcripts of previously acquired spoken interactions with 33 working professionals in pharmacy.

These 33 transcripts were collected during a previous study of the Interpersonal Simulator. The scenario in the previous study is the same as the scenario for the HDCM

model examined in the current evaluation (Vic Johnson with dyspepsia). The CCM model for that study was created by three knowledge engineers and two Pharmacy Experts using interactions with 51 pharmacy students. Creation of the CCM model required 6 months and approximately 200 hours. During the Interpersonal Simulator study, 35 participants interacted with a life-size VH using spoken inputs. Spoken inputs were recognized and transcribed using Dragon Naturally Speaking 9.5.

For comparative analysis, two of the 35 transcripts were removed from the test set due to speech understanding errors caused by accents. The transcripts from the remaining 33 interactions are used below to compare the accuracy of the conversational model created using HDCM to the model of the same scenario created using CCM.

3.3.1.2 Results

Conversational Modeling Time. Part 1 of the evaluation examined the progress of conversational modeling during the two weeks of development and how much total time was required of the domain expert. There were three iterations of conversational modeling improvement – group TA, group S1, and group S2. Participants interacted for an average of 20 minutes, making the total student time 62 hours. These three rounds of user-testing required 15 hours of expert time (including 2 hours of training time and 13 hours of suggestion processing and script editing) over a period of 2 weeks and created a conversational corpus consisting of 2655 stimuli and 595 responses, these results are summarized in Table 3-1 alongside the results for the previously created CCM model.

Table 3-1. Conversational Modeling Time Requirements for Centralized Conversational Modeling vs. Human-Centered Distributed Conversational Modeling

Method	Creators	Interactions	Expert Time	Novice Time	Stimuli	Responses
CCM	Knowledge Engineers, Pharmacy Experts, 51 Students	Spoken Interactions	~200 Hours (combined educator and knowledge engineers)	11 Hours (13 Minute Average)	1418	303
HDCM	Pharmacy Instructor & 186 Pharmacy Students	Virtual People Factory: Web-browser	15 Hours	62 Hours (20 minute average)	2655	595

Conversation Accuracy Improvements. Part 2 of the evaluation examined the trend of accuracy change for each group of participants. We evaluated the interaction transcripts for accuracy by reviewing the response to each participant question. We marked the response as accurate if there was a semantic link between the stimuli and response (Leuski et al., 2006); meaning there was a response and it was correct according to Vic's symptoms and medical history. We analyzed the percentage of responses that were accurate for all of group TA, and a simple random sample of 10 transcripts from groups S1 and S2. Figure 3-7 shows the percentage of responses that were accurate for all of group TA, and a random 10 transcripts from groups S1 and S2. The standard deviation of these samples is represented by the error bar in the figure, and exact standard deviations are provided in the caption. This analysis was performed only to establish a trend of increasing accuracy; the important accuracy is how well the conversational model performs with spoken inputs in comparison to the previously created CCM conversational model. The final accuracy analysis with spoken inputs is provided in the next subsection.

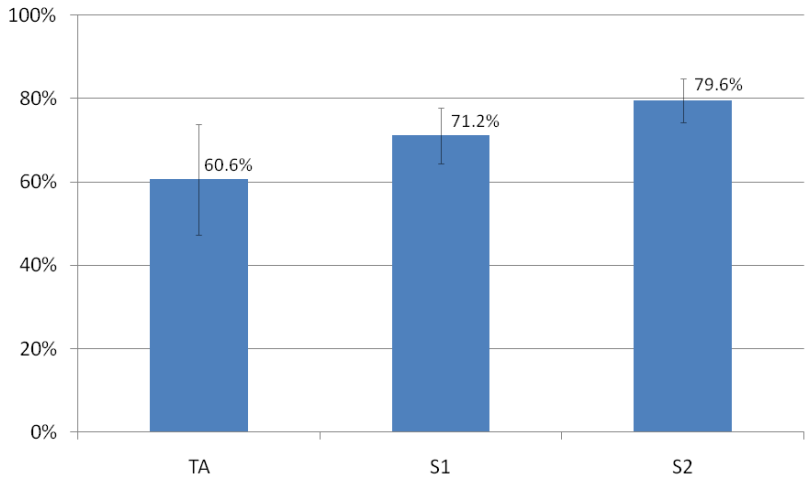

Figure 3-7. The accuracy of the dyspepsia conversational model for each group TA s.d. = 13.3%, S1 s.d. = 6.7%, S2 s.d. = 5.3%, represented by the error bars

Accuracy with Spoken Inputs. Part 3 of the evaluation compared the accuracy of the current HDCM conversational model to the accuracy of a previously created CCM model for spoken inputs. After the testing and improvements of the case study, we examined the performance of the HDCM model with spoken transcripts and compared that accuracy to the performance of a conversational model created using CCM. To run the comparison, we analyzed the transcripts from 33 spoken interactions between pharmacy students and the previous VH patient with dyspepsia.

During the interactions, Dragon Naturally Speaking 9.5 was able to transcribe spoken utterances at 83.3% accuracy. In order to analyze the spoken transcripts against the CCM and HDCM conversational models, we first removed inaccurate utterances due to speech recognition errors from the transcripts (16.7%). Utterances from the spoken transcripts were designated accurate if a human reader would have been capable of responding correctly. We then processed the remaining utterances using both the HDCM and CCM conversational models. Utterances were processed by feeding each utterance as a stimulus into a simulated conversation using each

conversational model. We then analyzed the accuracy of each response. Accuracy analysis revealed 74.5% accuracy (s.d. = 11.1%) per transcript for the conversational model created with CCM while the one created with HDCM had 78.6% accuracy (s.d. = 9.7%) per transcript for the 33 spoken transcripts. The accuracy data follows a normal distribution with a Shapiro-Wilk significance of .353 for CCM accuracy and .320 for HDCM accuracy (where values greater than .05 indicate a normal distribution). Samples for this analysis were paired because each transcript was processed using both conversational models. Using a paired samples T-test on the accuracy numbers for each transcript, we see a significant difference at $p < .05$ with $t = 2.4$. A summary of these results is shown in Figure 3-8.

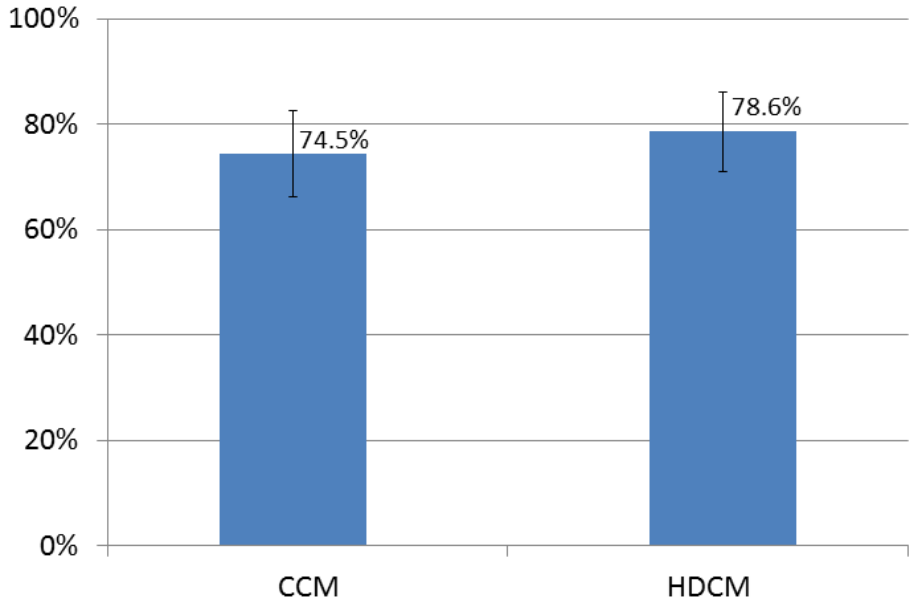

Figure 3-8. Accuracy of the Centralized Conversational Model vs. Human-centered Distributed Conversational Model for 33 spoken transcripts, improvement of 4.1% is significant at $p < .05$. CCM s.d. = 11.1%, HDCM s.d. = 9.7%, represented by the error bars.

3.3.1.3 Discussion

The results of this case study indicate that HDCM saves expert and developer time in creating the speech-understanding portion of a conversational model in

comparison to CCM. Using HDCM for conversational modeling yielded a significant 4.1% improvement for spoken interactions in ~7.5% of the expert time. Further, the conversation corpus created with HDCM has increased depth in the topics that students most frequently asked about. For example, there are only 44 questions about Aspirin in the corpus created with CCM, while there are 174 questions about Aspirin in the HDCM corpus. From this difference, we see that the pharmacy students concentrated on the medications the patient was taking, and HDCM led to a much larger number of medication related stimuli and responses, and thus allows a more nuanced conversation. Using HDCM, the pharmacy instructor was able to develop Vic in approximately 15 hours over 2 weeks, compared to the knowledge engineers and pharmacy experts creating Vic in ~200 hours over 6 months. Table 3-1 shows the differences in time input and conversational model output resulting from using the CCM and HDCM methods. We see that there is a decrease in the expert time by ~92.5% and increase in the total novice time by 545.5%. Involving this many novices in the conversational modeling process is possible because of the reduced logistical constraints provided by HDCM. Given such a large amount of novice data and an effective method for processing this data, the pharmacy instructor was able to create a corpus of nearly double the size of the CCM method.

The pharmacy instructor reported an additional advantage of the expert being directly involved; she was able to come up with new stimuli as she processed student stimuli. Often, a student's question would remind the pharmacy instructor of other stimuli and responses that should be in the conversation. The instructor would see a question such as, "How often do you take Aspirin?" and it would remind her that Vic

should also be able to answer if he is taking adult Aspirin, baby Aspirin, or Enteric Coated Aspirin. As a result, she would add new stimuli and responses so Vic could discuss those topics.

Feedback from both the pharmacy educator and pharmacy students stated that the experience was educationally beneficial. Surveys from student participants (groups S1 and S2) show “educational value” ratings of 49% positive (ratings 7-10), 28% neutral (ratings 5-6) and 23% negative (ratings 1-4). A breakdown of individual student ratings can be seen in Figure 3-9.

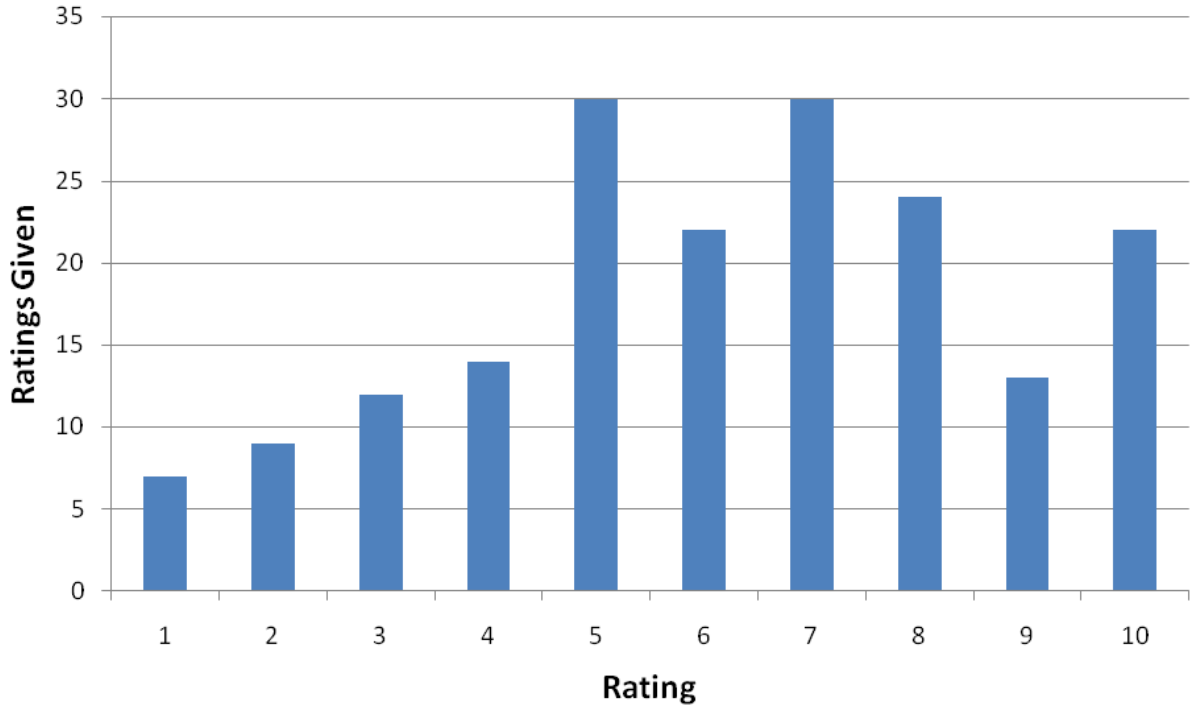

Figure 3-9. Student participant ratings on the educational value of interacting with the virtual patient.

The Pharmacy Instructor expressed that “building this scenario was relatively easy with minimal training, and that the effort is worthwhile because the scenario can be used over and over”.

3.3.2 Evaluation 2: Perceived Efficacy of Virtual People Factory for Healthcare Education

The success of VPF in the pharmacy domain prompted healthcare educators to use VPF in additional healthcare domains. Since the initial pharmacy study, healthcare educators have used VPF to create 23 additional scenarios used for teaching healthcare interview skills, we report on four of these scenarios. These four scenarios were chosen because they have been used for published research (Foster et al., 2010b; Shah et al., 2012; Shah et al., 2009a; Surkunalingam et al., 2009). Healthcare experts successfully created these conversational VHs using VPF with minimal phone and email assistance from knowledge engineers. Below we report on a meta-evaluation of the parts of their studies that pertained to VPF, in particular, the perceived educational value and usability of VPF.

3.3.2.1 Methods

In this evaluation, we analyzed a series of studies on the perceived efficacy of VPF VHs for healthcare education. The VHs were authored using VPF at the Medical College of Georgia (now named the Georgia Health Sciences University) and the Philadelphia College of Osteopathic Medicine. Because each of these studies was performed independently, their methods differ. We highlight aspects of each study that evaluate the efficacy of VPF for healthcare education from a student perspective.

Each of the studies received IRB approval before participants interacted with the virtual patient. Students participated in the studies on a voluntary basis, and the only inclusion criteria was enrollment in the surgical, psychiatry, pharmacy, or osteopathy program using VPF. Each of these educational interventions was used in addition to

traditional training methods such as standardized patient interactions and expert mentoring.

Surgical Patient with Melanoma. This study evaluated VPF's perceived educational benefit for medical education. Researchers at the Medical College of Georgia evaluated VPF for use in their medical curriculum (Shah et al., 2009a). With minimal phone and email assistance of knowledge engineers, the medical educators conceptualized the patient scenario, seeded the virtual patient conversational model, and then developed the virtual patient's knowledge base. The patient, Hank Lowry, is a 58 year-old male with suspicious skin lesions on his back, chest, and shoulder.

The medical educators followed the same process outlined for creating the pharmacy patient in evaluation 1. After receiving a lecture on obtaining patient history, 51 first-year medical students interviewed Mr. Lowry using VPF's browser-based interaction. After the interaction, participants completed a survey regarding the interview's educational benefit.

Psychiatry Patient with Depression. This study evaluated VPF's perceived educational benefit for first, second, and third-year psychiatry students. Psychiatry researchers at the Medical College of Georgia evaluated VPF for use in teaching and assessing history-taking skills for psychiatry interactions (Shah et al., 2012). The researchers created Cynthia Young, a 21 year-old female patient with chief complaint of insomnia and fatigue. The researchers ran two studies to evaluate the perceived educational benefit of this patient for psychiatry education.

The psychiatry researchers followed the same process outlined for creating the pharmacy patient. Participants included 71 first and second-year psychiatry students

and 67 third-year students. After the interaction, participants completed a subjective survey regarding the system's educational efficacy. Because of the varying levels of experience of the participants, this study illustrates the uses and limitations of VPF interactions. Specifically, the researchers compared the self-reported efficacy for first and second-year students to the efficacy for the third-year students.

Osteopathy Patient with Neurological Disorder. This study evaluated VPF's perceived educational benefit for osteopathic education. Researchers at the Philadelphia College of Osteopathic Medicine created a virtual patient for history taking before a neurological examination (Surkunalingam et al., 2009). The neurological patient, Nelson Sanjaya, is a 20 year-old male who is complaining of a suspicious headache, general malaise, fever, and nuchal rigidity. 46 second-year medical students at the Philadelphia College of Osteopathic Medicine used the VPF browser interface to interview Mr. Nelson. After completing the interview, students completed a survey regarding the self-reported educational value of the application.

Psychiatry Patient with Bipolar Disorder. This study compared participant performance between VPF browser-based interactions and spoken life-size interactions using the Interpersonal Simulator. The Bipolar Disorder character was created by a different type of domain expert, a peer-support specialist at the Medical College of Georgia (Foster et al., 2010b). The peer support specialist is a former patient with bipolar disorder. She wished to convey herself as a virtual patient in order to train psychiatry students to help their patients. To this end, she created herself and her husband as virtual humans. The patient character is a woman who presents with psychotic bipolar disorder and later that night develops a crisis. In part-1 of the bipolar

scenario, the participant interacts with the patient to assess her current state. In part-2, the participant interacts with the husband after the patient has a depressive episode and attempts to commit suicide. 25 third and fourth-year medical students interacted with the scenario -- 15 of the participants interacted using the instant message browser-based character and 10 interacted with the life-sized character by speaking. After the interactions, domain experts evaluated the completeness of the content elicited during the interviews.

3.3.2.2 Results

Details of the VHs described in the previous section are presented in Table 3-2.

Table 3-2. Conversational models created since the completion of the study.

Scenario	Name	Users	Stimuli	Responses	Modeling Time
Melanoma Patient	Hank Lowry	Surgery	621	189	21 hours
Depression Patient	Cynthia Young	Psychiatry	1314	345	15 hours
Bipolar Disorder Patient	Denise	Psychiatry	1605	220	11 hours
Meningitis Patient	Nelson Sanjaya	Osteopathy	777	228	25 hours

Hank Lowry, Surgical History Patient with Melanoma. After conversing with Hank Lowry, participants filled out a survey on the educational efficacy of the experience (all questions were on a 10 point scale, and 7-10 was considered positive), the results are reported in Table 3-3.

Table 3-3. Percentage of participants reporting 7-10 (good-excellent) on Surgical History Patient with Melanoma post-interaction survey

Question	First-Year (N=51)
How much did you enjoy this interaction?	65%
Do you feel this interaction was a valuable learning experience?	73%
How easy was it to use Virtual People Factory?	77%

Cynthia Young, Psychiatry Patient with Depression. After conversing with Cynthia Young, participants filled out a survey on the educational efficacy of the experience, the results are reported in Table 3-4.

Table 3-4. Percentage of participants reporting 4-5 (good-excellent) on Psychiatry Patient with Depression post-interaction survey

Question	First and Second-Year (N=71)	Third-Year (N=67)
Helped learn to formulate questions about depression symptoms	57%	31%
Valuable educational tool	66%	24%
Easy to use	71%	66%

Open-ended feedback suggested that this virtual patient experience is particularly useful in the first two years of medical school to decrease anxiety and offer practice before interviewing real patients during the clerkship years. Third-year students reported lower usefulness, and reported that the system would have been more useful in their first two years.

Nelson Sanjaya, Neurological Virtual Patient with Meningitis. After conversing with Nelson Sanjaya, participants filled out a survey on the educational efficacy of the experience, the results are reported in Table 3-5.

Table 3-5. Percentage of participants reporting 4-5 (good-excellent) on Neurological Patient with Meningitis post-interaction survey

Question	Second-Year (N=46)
Beneficial in preparation for live patient encounters	79%
Valuable educational tool	71%
User-friendly	71%
Would like to have VPF virtual patients available for future training	92%

Denise, Psychiatry Patient with Bipolar Disorder. This study compared browser-based interactions to spoken life-size interactions. The two systems were compared based on the content elicited during the interaction as seen in Table 3-6.

Participants interacting with the virtual patient using a browser were more likely to ask about suicide, grandiosity, and elevated mood; while participants interacting with the virtual patient in the life-size interaction were more likely to ask about distractibility and the duration of the illness. From this study, we see that 1) students can successfully perform an assessment in either medium and 2) the differences in the medium may cause the users to focus on different topics in the interaction.

Table 3-6. Comparison of content elicited in browser versus Interpersonal Simulator interviews with Psychiatry Patient having Bipolar Disorder

	Browser (N=15)	Interpersonal Simulator (N=10)
Suicidal Ideation	100%	80%
Grandiosity	73%	40%
Elevated mood	93%	80%
Distractibility	60%	80%
Illness duration	60%	100%

3.3.2.3 Discussion

These studies indicate that VPF may be a viable and well-received method for augmenting current interview training curricula. The domain experts who ran these studies remarked that VPF provides an alternative method for practicing patient interviews in a resource-, time-, and cost-effective manner. In a prior study of US and Canadian medical schools, 74% took 3 months to more than 2 years full-time to develop a single virtual patient scenario, compare this to the 11-25 hours of Table 3-2 to develop a virtual patient using VPF (Huang et al., 2007). The medical educators further stated that VPF allowed medical students to learn correct history taking techniques prior to interacting with patients in the clinic.

Students who used these scenarios have remarked “Great tool to practice taking history.” and “I think that this is a great program… As first and second-year students,

interviewing patients can be very nerve wracking and this may be a great bridge to becoming relaxed in patient interviewing." We find that VPF based virtual patients are particularly useful during the first two years, but have decreased utility as the students reach the third-year and beyond. This may be because students begin frequent interactions with both standardized patients and real patients during their third-year. This finding is related to the types of scenarios involved in these studies; given more advanced topics we may find a different trend.

The majority of students found their interactions to be educationally valuable if placed in an appropriate stage of the curriculum. Further, the educators were able to create educational virtual patients themselves in collaboration with students, and with minimal assistance from computer scientists. Medical educators have indicated these properties are essential in order to see widespread adoption of virtual patients in the healthcare field.

3.3.3 Evaluation 3: Usability for healthcare education

The final evaluation assesses domain expert feedback on the usability and acceptability of VPF and HDCM. Usability and acceptability are measured using the following metrics:

1 **Usability:** the domain expert's perceived ease of use
2 **Acceptability:** the domain expert's perceived usefulness for education

3.3.3.1 Methods

Domain experts (N=11) from the above research studies were issued a digital survey on usability and acceptability. This questionnaire assessed the domain expert's self-reported usability of VPF for healthcare education. It also assessed their self-reported view on how useful VPF is in preparing students for patient interviews. The

usability and acceptance survey was based on Davis' *Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology* survey (Davis, 1989). Davis' survey has been previously shown to have a high coefficient of reliability with Cronbach alpha of .98 for perceived usefulness and .94 for perceived ease of use. This questionnaire has 12 questions; responses were rated from 1 - unlikely, to 7 - likely. Questions were modified to refer specifically to VPF and medical education, for example: "I would find **the system** useful in my job" was changed to "I would find **Virtual People Factory** useful in **student education**". Three additional questions were added at the end of the survey to also assess if the educators found VPF to be educationally valuable, if they plan to continue using VPF in their courses, and if VPF could save time in teaching particular topics.

3.3.3.2 Results

Seven out of the eleven domain experts responded to the survey. Survey results indicate that, on average, domain experts felt the system was easy to use and would be useful for healthcare education, as seen in Table 3-7. Additionally, 3 out of 7 domain experts responded that they will continue to use VPF in their courses (4 responded that this option was not applicable as they were not teaching courses). Of the 3 domain experts who will continue to use VPF in teaching their courses, 2 out of 3 will use VPF to replace a portion of their lecture, and they note that this will save them on average 90 minutes of lecture time.

Further, to introduce a patient case, experts normally take 8-10 hours to create the paper and pencil case. If we consider that a VPF virtual patient takes 15 hours, those 5 extra hours allow the expert to provide an enhanced learning experience and the ability to distribute learning to students prior to coming to class. Prior distribution allows

students to be ready to discuss the case and be tested during the class. These advantages indicate that not only will the VPF interactions provide enhanced learning experiences – they can also be used to review patient cases that will then be discussed in the lecture, thus saving the time used to introduce the patient case and get students to the point where they can discuss the case.

Table 3-7. Results of the domain expert survey on Educational Value, Usefulness, and Ease of Use of Virtual People Factory. Responses were rated from 1 - unlikely, to 7 - likely.

	N	Mean	Std. Deviation
Educational Value	7	6	0.0
Usefulness			
Work more quickly	6	5.2	1.0
Job performance	7	5.9	0.4
Increase productivity	7	6.0	0.6
Effectiveness	7	6.0	0.8
Make job easier	4	5.5	0.6
Useful	7	6.4	0.5
Ease of Use			
Easy to Learn	7	6.1	0.9
Controllable	7	5.6	1.0
Clear & Understandable	7	5.6	0.5
Flexible	7	5.6	0.5
Easy to Become Skillful	7	6.1	1.2
Easy to Use	7	5.9	0.9

In discussing VPF, healthcare experts made the following negative remarks regarding time and accuracy:

- "Initial script-development process is time-consuming."
- "[Training the patient] is time consuming."
- "Interactions still elicit some inaccurate responses from the virtual patient."
- "[Even after] training for an answer in all the possible ways it can be asked it sometimes still doesn't respond properly."
- "It would be nice to be able to save questions as a template and plug in new answers based on the scenario."

In discussing VPF, healthcare experts made the following negative remarks regarding the web-browser interactions:

- “[VPF has a] limited ability to offer animation.”
- “Need greater interactivity.”
- “Need to be able to show facial expression for psych scripts.”
- “It can be difficult to get students to treat a VPF interaction as an actual clinical interaction.”

In discussing VPF, healthcare experts made the following positive remarks regarding ease of use:

- “[VPF is] easy to use once you learn the system involved in script development and analysis.”
- “I got caught up in creating the characters. It was fun to imagine this whole patient from scratch and then give them a history from their favorite food to their feelings of stress and depression related to office politics.”

In discussing VPF, healthcare experts made the following positive remarks regarding educational value:

- “[VPF interactions are] great beginning experience for medical students.”
- “[VPF is] very useful for helping students ask about sensitive issues.”
- “[VPF interactions are] an excellent way to review criteria for depression and bipolar disorder”
- “The experiential learning aspect of this program empowers students and gives them the skills to continue that learning in other aspects of their education.”
- “[Students] have the opportunity to review their own transcripts which is very useful and it makes people aware of areas of possible improvement”
- “With enough clinical scenarios, I feel VPF can become extremely valuable.”
- “I would love to see [VPF]’s development continue and become adopted in multiple places.”

3.3.3.3 Discussion

The results of evaluation 3 indicate that medical educators perceive VPF to be highly useful and beneficial to their students. The results further suggest that the majority of the healthcare educators will continue to use VPF in their courses. In courses, VPF will be used in conjunction with both lectures and standardized patient interactions. When combined with lectures, VPF interactions may be used to replace the initial discussion of a patient case, and then used to prompt additional discussion after the interactions. Some experts also state that they would recommend VPF for adoption.

Some negative remarks made by the experts include the process still being time consuming, that the VH patients still give inaccurate responses, and that the browser-based interactions are limited. In Chapter 4, we discuss Conversational Knowledge Reuse, which further reduces the time required and improves the accuracy of interactions. The browser-based interactions are limited by design, and it is understandable that students would not treat these interactions the same as a full clinical interaction. The intention is for VPF to be used for creating conversational models and for simple practice, and then full training interactions will be conducted using more immersive interfaces, such as the Interpersonal Simulator (Johnsen, 2008). While these limitations are present, the time and accuracy issues will continue to be reduced by future work, and the interactivity can be improved by using more immersive clients.

One limitation of this evaluation is that the educators involved with these studies are collaborators of the researchers and have put time into learning and using the system. There is an inherent desirability bias in the educator's responses to the self-

report survey and the sunk cost of learning the system may cause additional positive feedback influences.

These self-reported results are an initial check that educators are not only using VPF in healthcare education, but they also feel it is providing educational benefit. We further find that the paradigm of using HDCM to create educational virtual patients is acceptable to healthcare educators. This is an important finding as it is unusual to ask students to be a part of the creation of educational materials. Generally, educators create materials on their own and then use those materials in class. These initial results indicate that involving students as active participants in the creation process may be a successful technique.

CHAPTER 4
CONVERSATIONAL KNOWLEDGE REUSE: VIRTUAL HUMAN TEMPLATES AND DYNAMIC KNOWLEDGE SHARING

This chapter presents the analysis, design, implementation, and evaluation of two new Conversational Knowledge Reuse methods, *Virtual Human Templates* and *Dynamic Knowledge Sharing* (research questions 1-3). These two methods provide new sources of conversational knowledge and help further reduce the time taken for conversational modeling of virtual humans.

We describe a study which used *Virtual Human Templates* and *Dynamic Knowledge Sharing* along with Human-centered Distributed Conversational Modeling to generate seven new virtual human patients within the context of a healthcare course. Students in the course (N=32) authored the virtual human patients themselves. Integrating the creation of virtual human patients into a course was previously infeasible due to time requirements, and was made possible by the addition of *Virtual Human Templates* and *Dynamic Knowledge Sharing* to the conversational modeling process. This study provides evidence for the usability and efficiency of Conversational Knowledge Reuse in the creation of a diverse set of virtual humans. A publication on the evaluation of these methods and implementations was submitted to the Journal of Autonomous Agents and Multi-Agent Systems and is currently in review (impact factor = 2.1) (Rossen et al., 2012).

Personal Contributions. I conceptualized the Conversational Knowledge Reuse concept and the *Virtual Human Templates* and *Dynamic Knowledge Sharing* methods, implemented the *Virtual Human Templates* system, extended Virtual People Factory to use *Dynamic Knowledge Sharing,* and designed, ran, and analyzed the study.

Collaborators. Shivashankar Halan was involved in the design of the *Virtual Human Templates* system and assembled the template used in the study. He was also involved in the design and implementation of the study. Dr. Michael Crary provided access to study participants (dysphagia students) and provided valuable feedback in the design and evaluation of the study.

Relevance to thesis. This chapter describes two new methods and their integration into the Virtual People Factory system. These additions are evaluated in a user study which demonstrates a reduction in the time to model virtual human conversations. This evaluation also shows that these methods make the process feasible for domain novices to rapidly generate new virtual human conversational agents within the context of a healthcare course.

4.1 Overview

Applications of virtual human (VH) interpersonal skills training require a diverse set of VH scenarios. Interviewing skills cannot be trained by interviewing the same person over and over; learning these skills requires diverse experiences covering many issues (Huang et al., 2007). Providing diverse experiences is perhaps the most compelling reason to develop VHs for interpersonal skills education (Johnsen, 2008).

VHs can be modeled to provide a wide variety of human aesthetic characteristics in combination with a variety of issues to provide both contextually and ethnically diverse experiences (Rossen et al., 2008; Rossen and Lok, 2012). VHs can also simulate scenarios that students may otherwise never encounter in training or that take place in a setting inappropriate for training (Kotranza et al., 2009a). However, using previous technology, acquiring the knowledge necessary for each VH simulation has been costly in terms of both time and money (Dickerson et al., 2005; Glass et al., 2005;

Huang et al., 2007; Kenny et al., 2007; Kenny et al., 2008; Villaume et al., 2006). To address the limitations of conversational modeling, we propose a novel method of generating new conversational models from existing conversational data, Conversational Knowledge Reuse.

The Conversational Knowledge Reuse method allows VH creators to leverage previous and ongoing efforts towards creating similar VHs. The concept is that small portions of conversational data, such as sentences in a transcript or question-response pairs from a conversational model, can be processed and reused to generate new unique conversational models. For example, we can create the conversational model for a virtual patient with a stomach ulcer, and then reuse portions of that conversational model to create a virtual patient with appendicitis. The Conversational Knowledge Reuse method will enable VH authors to create a wider variety of VHs, and thereby provide more diverse experiences using VHs.

Chapter 3 described work on conversational modeling using Human-centered Distributed Conversational Modeling (HDCM). Using Conversational Knowledge Reuse, we extend the conversational modeling process defined in HDCM to include new methods of reusing conversational data for the creation of VH conversational models. Employing Conversational Knowledge Reuse results in robust conversational models in significantly shorter time. This rapidity opens up new application areas such as producing diverse sets of VHs.

In this chapter we explore two implementations of Conversational Knowledge Reuse 1) *Virtual Human Templates* (VHT) and 2) *Dynamic Knowledge Sharing* (DKS). These two applications of Conversational Knowledge Reuse share the underlying

method of reusing conversational data acquired for the creation of one conversational simulation in the creation of a different conversational simulation. The process for VHT extracts the knowledge from a previously created conversational model and stores it in a template. To create a template, a knowledge engineer processes conversation specific knowledge into knowledge that is applicable to a whole domain of conversations, such as patient interviews. This generalized information is then customized by a VH author to generate a new conversational model (details in section 4.2). In contrast to the preprocessed knowledge from VHT, DKS knowledge is reused dynamically and shared among a group of similar VHs during creation. Because the VHs are similar, the VH conversational models for each scenario should all be able to respond correctly to the stimuli for any of the other scenarios. This stimulus knowledge is shared by simulating interactions from one VH's conversations to all of the other VHs in the group, and thereby discovering what questions the other VHs cannot yet respond to correctly (details in section 4.3).

Using the methods of Conversational Knowledge Reuse, the authors of VH conversational models do not need to start over every time they create a new VH. Further, the process of creating VH templates develops a warehouse of knowledge to draw from. Using this warehouse, each successive VH conversational model is more robust and faster to create than the previous generation of VHs.

4.2 Virtual Human Templates

This section describes the creation of a virtual human template, the generation of a new conversational model using a template, and the process of refining that model after it is generated. A knowledge engineer processes the original conversational model to turn it into a virtual human template (section 4.2.1). A virtual human template is a

generalized corpus of knowledge. VHT uses the template to construct a Virtual Patient Generator interface. Using the Virtual Patient Generator, a VH author fills in the parameters of the template, and the system generates a conversational model customized to their particular scenario. The generated conversational model is compatible with Virtual People Factory, and is now ready for the HDCM process. An overview of the process for creating and using a virtual human template is shown in Figure 4-1.

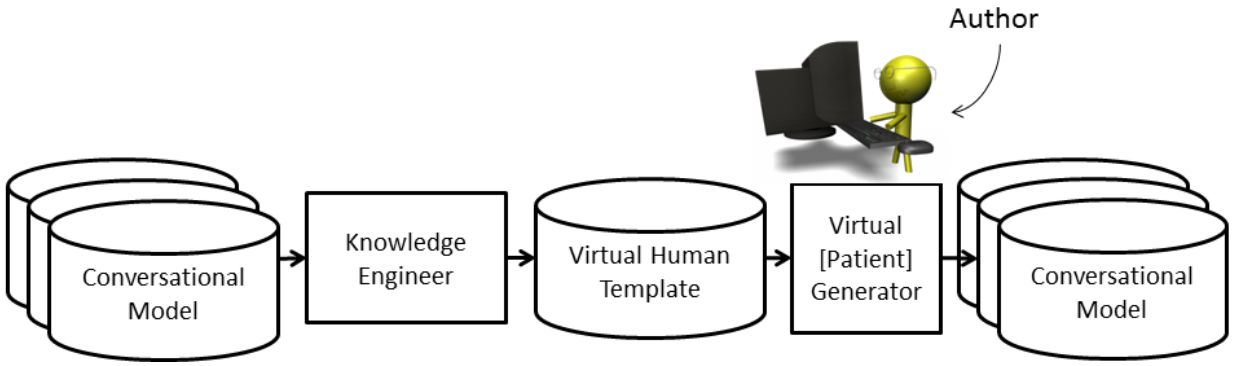

Figure 4-1. The construction and use of a virtual human template to construct conversational models in the *Virtual Human Templates* system

4.2.1 Virtual Human Template Creation

A virtual human template is created based on one or several existing conversational models. The corpora of the original conversational models are reviewed and generalized by a knowledge engineer. A knowledge engineer generates a virtual human template by reviewing each stimulus and processing the generalizable stimuli and responses from the original conversational model(s). To process a single stimulus and response, the knowledge engineer first examines the set of similar stimuli and responses. The overlapping information that is specific to the original conversational model's scenario is then marked as a variable. This parameterized stimulus and response are then inserted into the template.

A small example template presenting several stimulus-response pairs is shown Table 4-1. Within the template, the scenario specific information from the original responses “Vic Johnson”, “110/80”, and “170” are turned into [var]s for customization. The resulting template facilitates the generation of new conversational models, as described in section 4.2.2.

Table 4-1. Example virtual human patient template

Original Stimulus	**Original Response**	**Template Stimulus**	**Template Response**
Are you Vic Johnson?	I am Vic Johnson	Are you [var]?	I am [var]
What is your blood pressure?	My blood pressure is 110/80	What is your blood pressure?	My blood pressure is [var]
Any family history of diseases?	Diabetes runs in my family	Any family history of diseases?	[var] runs in my family

4.2.2 Conversational Model Generation

The VH template created by the knowledge engineer is fed into a “virtual X generator”. For the template used in the current study, X = “patient”. Authors of VH patients use the “virtual patient generator” to customize the template and generate new VH conversational models. A screenshot of the interface for the virtual patient generator can be seen in Figure 4-2. As shown in the figure, VH authors fill in the [var] information for each question, and the generator uses that information to create the customized conversational model. For example, when the author fills in “Diabetes” in the text field of the example (Figure 4-2), the generator customizes the response “[var] runs in my family” to “**Diabetes** runs in my family”. It also customizes any occurrences of [var] in the 23 paired stimuli that are linked to this response. As an alternative to filling in the [var], the VH author may choose to completely replace the response by pressing

"Replace response", and instead of saying "Diabetes runs in my family", the VH could say "My mother and grandmother both had diabetes".

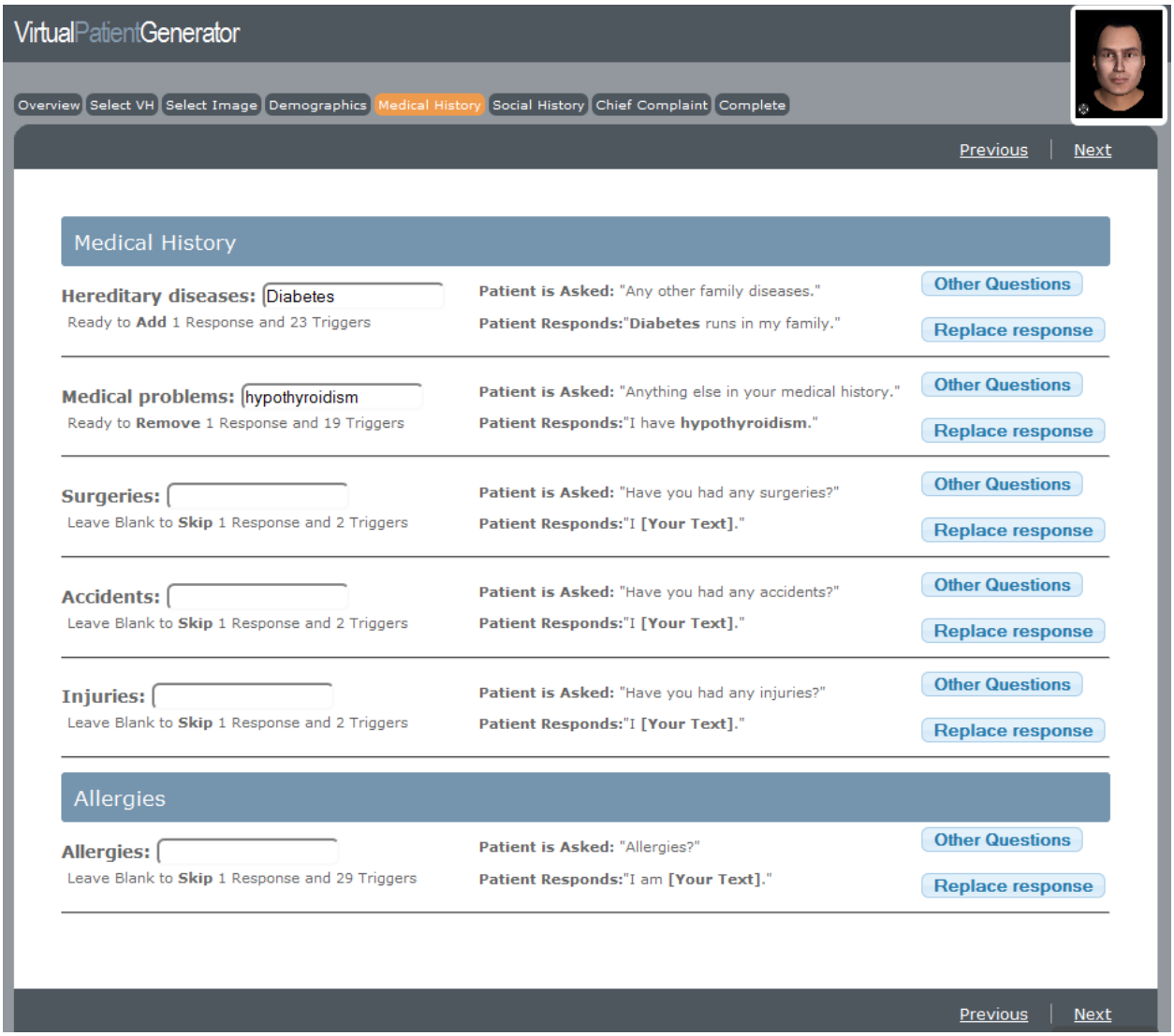

Figure 4-2. A screenshot of the Virtual Patient Generator used in the study. The Virtual Patient Generator interface was programmatically constructed by the Virtual Human Templates system based on the virtual patient template.

The process of using the virtual patient generator currently consists of five steps, the last of which transfers the new conversational model to Virtual People Factory. Within Virtual People Factory, the conversational model is refined using the HDCM process described in Chapter 3. The following are the steps of virtual patient generation:

1. Overview: Provides a brief description of the upcoming process.

2. Select VH: Initiates the creation of a new VH or allows the selection of a previously worked on VH.

3. Select Image: Allows the selection of an image for the VH. The selection includes a range of faces as shown in Figure 4-3. These faces were created using Autodesk's Evolver.

4. Customize Template: In this step the author fills in the conversation specific information.

    a. Demographics: This is the first template based page. It includes template entries for the VH's name, race, weight, and gender.

    b. Medical History: Includes entries for hereditary diseases, existing medical problems, previous surgeries, accidents, and allergies.

    c. Social History: Includes entries for marital status and socially based risk factors. Risk factors include alcohol consumption, tobacco consumption, sexual history, and occupation.

    d. Chief Complaint: Includes entries regarding the chief complaint, history of the complaint, and symptoms associated with the complaint.

5. Complete: This is the final step in the patient generation process. The user confirms their choices and the system generates a conversational model from the customized template. The system then transfers that model to Virtual People Factory.

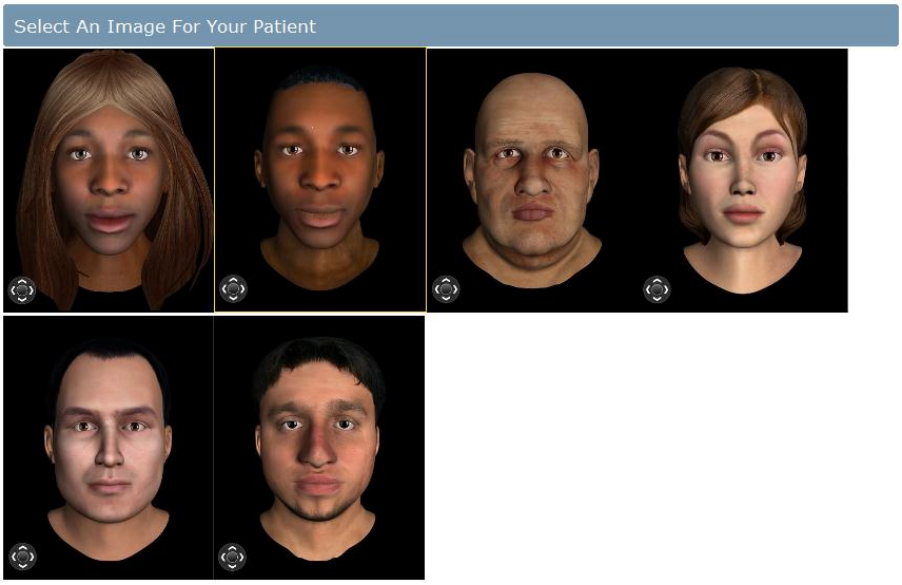

Figure 4-3. The faces available for selection in the Virtual Patient Generator

4.2.3 Conversational Model Refinement

After the VHT system transfers the conversational model to Virtual People Factory, the author is redirected to Virtual People Factory to continue developing the conversational model using the HDCM process. The template has seeded the process of HDCM, which will speed up the conversational modeling, but VHT does not supersede HDCM. Templates are not used in isolation because they cannot completely cover the stimuli-response space of a particular conversation. We have previously discussed in Chapter 3 that we cannot predict the complete stimuli-response space of a conversation. VH templates cover only the overlapping information of an interaction domain, which in this case is patient interviewing. The template covers common patient interaction stimuli such as “What is your name”, “What medications are you taking”, and “What brings you to the office today”. These questions are common in many patient interactions, and therefore cover the first layer of stimuli that are likely to be encountered. However, the template may not cover the stimuli and responses to follow up questions. For example, if the response to “What medications are you taking” is “I’m taking Zestril and Synthroid”, and the follow up question is, “Do you take Zestril every day”, the template may not cover that stimulus because it is specific to the current scenario. The *Virtual Human Templates* system is capable of covering this additional question, but will only do so if it was covered in the source conversational model. For these reasons, creating a VH conversational model using a virtual human template still requires HDCM to prepare the model for high quality VH interactions.

4.3 Dynamic Knowledge Sharing

The second method of Conversational Knowledge Reuse explored is DKS. In contrast to VHT, which statically leverages conversational model overlap, DKS

dynamically leverages that same overlap. VHT is static because it requires a knowledge engineer to process existing knowledge and put it into a statically stored format. DKS is dynamic because it reuses knowledge immediately as it appears in the system. The advantage of VHT is that the knowledge has been validated and converted into a format that can be used rapidly, but there is a delay before the knowledge is available. Using the DKS method, knowledge can be made available immediately, but must be processed by the VH author in the same way HDCM knowledge is processed.

DKS is used when a group of similar VH conversational models are being developed at the same time. The knowledge from these scenarios may be shared among the conversational models to improve the total knowledge available for modeling each conversation. The way DKS shares this knowledge is by taking each conversation conducted with one of the VHs and simulating that conversation with all of the other VHs. DKS allows stimuli from interactions with one VH to be tested against the current knowledge of the other VHs, and any new stimuli are added into the *Suggestions System* (described in Chapter 3). An overview of the DKS process can be seen in Figure 4-4. In the DKS process, end-users (who are generally novices) conduct an interaction with one of the VHs using its own conversational model. That same interaction is then simulated with each of the other conversational models. Simulation consists of testing each stimulus from the interaction against each of the other conversational models, as though the interaction was occurring with each of the VHs individually. Whenever a VH cannot respond to a stimulus, that stimulus is fed into that VH's suggestions list. The VH author processes these suggestions using the suggestions system, the same way they are processed in HDCM.

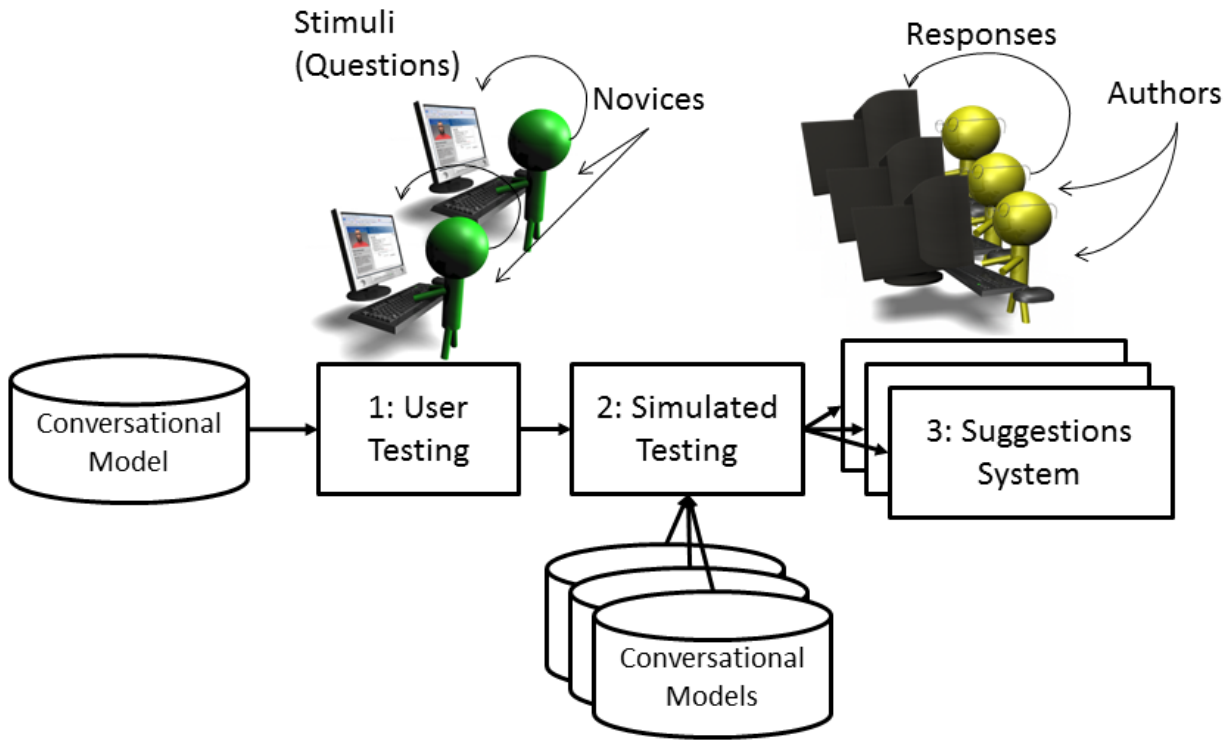

Figure 4-4. Overview of the Dynamic Knowledge Sharing process. 1: Novice users interact with the VH using its conversational model; 2: the system simulates that interaction with other VH's using their conversational models and unknown stimuli are added to the simulated VH's Suggestions System; 3: the unknown stimuli are validated by VH authors, paired with a response, and placed in the conversational model.

As can be anticipated, because these stimuli were acquired during conversations with a different VH, the relevance of the stimuli acquired using DKS is often lower than the relevance of conversations intended for that specific VH. However, the net effect is an increased usage of knowledge acquired from every novice interaction, which results in more total knowledge for all of the conversational models. The relevance of stimuli acquired from these resources is analyzed in section 4.4.4 and discussed in section 4.4.5.

4.4 Study

4.4.1 Methods

The Conversational Knowledge Reuse methods were evaluated within a course. The authors of the VHs were the students of that course. To the best of our knowledge, having students generate natural language conversational models has not been tried before and was previously infeasible due to the time required to create VH conversational models. Having students create multiple choice virtual patients was previously explored and was found to be overly challenging for the students (Villaume et al., 2006).

The experiment was conducted within a course called Dysphagia Management. Dysphagia is the medical term for the symptom of difficulty swallowing and can be observed in patients suffering from a variety of medical conditions. The course was taught over a period of four months (one semester) during the spring of 2011.

Students in the Dysphagia Management course are preparing to be clinical practitioners in the field of Speech-Language Pathology. After these students graduate, patient interviews will be critical to the quality of their clinical practice. Traditionally, the only way these students learn patient interviewing skills is by practicing with real patients under the supervision of a professor. Since a professor is required for supervising practice interview experiences, students are provided with a limited number of practice opportunities. The Dysphagia Management course prepares students on the theory of patient interviewing for dysphagia related conditions, but does not provide them with real-patient interactions. As such, the course instructor felt the students would receive benefit from creating virtual patients and using those patients to practice their interviewing skills.

To implement the proposed Conversational Knowledge Reuse methodology within the course, students worked in small groups to create and revise their own virtual patient in scheduled exercises throughout the semester. During creation of the conversational models, the students were not required to use a particular conversational modeling method, and were instructed in how to use all four methods: Manual, HDCM, VHT, and DKS.

Manual. Manual additions are the direct method of entering information into a conversational model. The author uses the Virtual People Factory Edit Scripts interface to manually add new stimuli and connect them to new responses.

HDCM. HDCM knowledge comes from either interaction between a public user and a VH, or between a VH author and their own VH. Stimuli from public interactions are collected automatically during interactions with public interviewers (non-author interviewers) when the VH does not respond, or when the user manually identifies that the VH has responded incorrectly. Stimuli from author interactions are collected similarly to public interactions, except the interviewers are the VH authors. Authors interview their own VH to "try out" the interaction and either the VH does not respond or the author manually identifies that the VH has responded incorrectly (additional details in Chapter 3). Knowledge from this method is pushed into the *suggestions system*, where it is processed to become part of the conversational model.

Virtual Human Templates. Stimuli from this source come from the virtual patient generator during the customization of the template for the initial seeding of the conversational model.

Dynamic Knowledge Sharing. Stimuli from this source come from the DKS system and are similar to a public true-negative, but are collected from interactions with other conversational VHs. Questions the conversational model does not contain are pushed into the *suggestions system*, where they are processed to become part of the conversational model.

For the purposes of evaluating Conversational Knowledge Reuse, the data analysis focuses on 1) evaluating which modeling methods the students chose to use, 2) the accuracy of the resulting conversations and how that relates to the methods used, and 3) the time that went into each method and which strategies were most effective. Using this information, we established the relative effectiveness of Manual, HDCM, VHT, and DKS.

4.4.2 Population

The course consisted of 32 health profession students working towards a graduate degree in Speech-Language Pathology. All 32 of the students in the course chose to participate in the study. The class was predominantly female (87.5%). Participants were first- or second-year graduate students (72% first-year).

4.4.3 Procedure

In the first week of class in spring 2011, students received lectures on the basic causes of dysphagia. Students were then introduced to Virtual People Factory by the experimenters during a class room lecture. Before starting the study, participants received and signed an IRB consent form. The students were told that participation in the study was optional, and that an alternative assignment would be provided if they chose not to participate. All students in the course chose to participate. The study proceeded as shown in Figure 4-5.

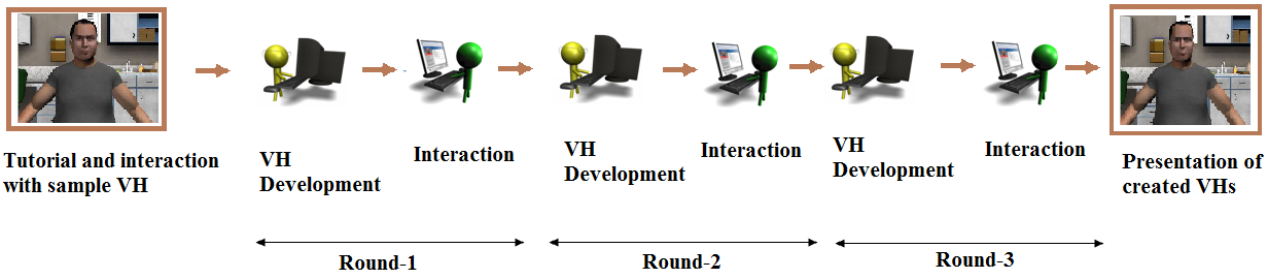

Figure 4-5. Conversational Knowledge Reuse evaluation study procedure

After signing the consent form, participants received a one-hour introductory training on Virtual People Factory. As part of the training, the experimenters had students practice interviewing Vic Johnson, a patient created in a prior pharmacy class, described in Chapter 3. They were then issued individual logins for Virtual People Factory and given their group assignments. The participants were randomly assigned into seven groups of four or five. Each group created one VH patient.

Each group was provided with a profile for a patient to produce. The profile (created by the course instructor) included a brief background of the patient containing gender and family history and a description of the patient's dysphagia condition. Participants made up the name, culture, and personality of the patient. The patient scenarios were all similar in that the patient presented with a swallowing disorder. However, one of the seven was conversationally dissimilar because the patient was a baby, and the conversation was conducted with the child's mother. Due to this difference, this conversation is used as a contrasting example in the data analysis. The seven scenarios are shown in Table 4-2.

Participants constructed their VH outside of class time, and the amount of time spent on their VH was strictly voluntary. Virtual People Factory tracked the amount of

time each participant spent creating their VH. Participants were aware that this time was being recorded.

Table 4-2. Virtual Humans created as part of the experimental study. ‡Scenario in which the patient is the baby and the VH is the mother.

Name of virtual patient	Number of Students	Diagnosis	Culture / Sex	Personality
Marty Graw	4	Esophageal Stricture	Haitian / Male	Anxious and amiable
Vinny Devito	5	Brainstem Stroke	Italian-American / Male	Loves food and spending time with family
Jackie Dauer	5	Supraglottic Laryngectomy	African-American / Female	62-year old widow who is friendly
Kahlua Lopez	5	Left Hemisphere Stroke	Hawaiian / Female	Lower middle class (English is second language)
Johnny A Seed	4	Zenker's Diverticulum	Caucasian / Male	Elderly male, Lively, happy, optimistic
John Smith	4	Head and Neck Cancer with Radiation	None / Male	None
Anne Animus‡	5	Baby has GERD	Caucasian / Female	Single Mother

Participants used a virtual patient template to create their initial VH conversational model. To refine the content the VH was able to discuss they used the HDCM with DKS process as well as manually entering new stimuli and responses. Participants refined their conversational model using three rounds of interleaved VH development and patient interview interactions (as shown in Figure 4-5). During a round, each participant was asked to interact individually with their virtual patient, and then interact individually with two other teams' virtual patients. By the end of the third round, every participant was asked to interacted with everyone else's patient (six interactions), and interact with their own patient three times. A visual representation of the study procedure is shown in

Figure 4-5. At the last week of class, the instructor conducted a review of the created virtual patients in class, as seen in Figure 4-6.

Figure 4-6. Student groups presenting VHs they created to the class and instructor

4.4.4 Data Analysis

We performed an analysis of the sources of stimuli (Manual, HDCM, VHT, and DKS) contained in each conversational model. The analysis established the relative *usage, quality,* and *efficiency* of each conversational modeling method.

- *Usage (model)*: The percentage of each conversational model that originated from each method.
- *Usage (responses)*: The percentage of each transcript that originated from each method.
- *Quality*: The percentage of accurate responses during interactions originating from each method.
- *Efficiency (model)*: The increase in quality per hour spent on each conversational model.
- *Efficiency (method)*: The average number of stimuli added to a conversational model for each hour spent on a method.

To establish the usage, quality, and efficiency for each of the methods, we computed the following metrics for each method: percentage of each conversational model, percentage used in interactions, accuracy of responses, and time spent. These metrics were also calculated for each conversational model as a whole to establish the effect the methods had on the overall quality of the conversational model. The remainder of this section describes each of the metrics.

Percentage of the conversational model. We calculated the percentage of the conversational models that originated from each of the methods. The percentage of stimuli sources that made up each conversational model shows the usage of each method to build the conversational model, but does not show if these stimuli were used during interactions.

Percentage used in interactions. During an interaction between a human and a VH, each utterance from the human may cause a response from the VH. Utterances are either questions or statements typed by the human interviewer. We calculated the percentage of responses that originated from each method. The percentage of responses from each method during interactions shows what percentage of the actual conversations originated from each method. Usage was analyzed using the final (3rd) round of transcripts because that was when the conversational models were most developed. Only transcripts of interactions with public users were analyzed for the usage of each source. By comparing the usage of each source with the *quality* of the interactions, we can evaluate how the method sources chosen by authors affected the performance of the conversational model in practice.

Accuracy of responses. While the distribution of response sources indicates the *usage* of each conversational source during interactions, it does not indicate the *quality* of those responses. We define *quality* as the percentage of *accurate* responses from each method. The *quality* portion of the analysis was determined by the accuracy of responses during interactions. Accuracies were also analyzed using the final (3rd) round of public interaction transcripts. To determine the accuracy, the transcripts were manually reviewed for the accuracy of the given responses. As in the analysis of HDCM (Chapter 3), the responses were marked as accurate if there was a semantic link between the utterance from the user and the response from the VH (Leuski et al., 2006), meaning there was a response and it was correct according to the VH patient's symptoms and medical history. Utterances that were problematic due to incompletely or incorrectly typed statements were removed from the transcripts prior to analysis.

Time spent. Time was calculated in hours spent on each Virtual People Factory page. Pages were grouped by conversational modeling method. Time spent idle on the homepage and time in which participants spent more than one and a half hours on a single page were considered periods of inactivity, and so were discounted from the time calculations. Overlapping time, in which multiple pages were open, were recorded as only the first page opened plus non-overlapping times.

The time spent on both HDCM and DKS together is the time authors spent using the Suggestions System. Because any time spent on the Suggestions System is spent processing suggestions from both HDCM and DKS, we computed the time based on the ratio of the number of suggestions processed for HDCM or DKS over the total number of suggestions processed.

4.4.5 Results

We first present an overview of the resulting conversational models, the time spent on those models, and the resulting accuracy of responses during interactions (Table 4-3). There were seven conversational models produced. To determine the accuracy, the transcripts from the final (3rd) round of interactions were manually reviewed for the accuracy of the given responses. The 3rd round of transcripts consisted of 51 transcripts with 6 to 9 interactions per VH. In those 51 transcripts there were 2365 utterances for an average of 46.37 utterances per transcript with s.d. = 25.14. The accuracy resulting from this analysis was used to sort the tables in this section.

Based on the accuracy results shown in Table 4-3, the VHs did not respond or responded incorrectly to between 10.28 and 16.44 utterances per interaction. The accuracy of the Anne Animus conversational model is lower than any of the others. The data analysis from the Anne Animus conversational model is included in the analyses to provide contrast to the other six models. The authors of the model were unable to use the Conversational Knowledge Reuse methods due to the dissimilarity of the scenario.

Table 4-3. The time required to create each virtual human, and the size of each virtual human conversational model, ordered by the accuracy of the models. ‡Scenario in which the patient is the baby and the VH is the mother. Anne Animus is shown for illustration purposes and not included in mean or s.d.

Name of virtual patient	Stimuli	Responses	Time (hours)	Accuracy
Marty Graw	1238	362	17.49	77.83%
Vinny Devito	1358	465	22.84	72.84%
Jackie Dauer	1576	557	14.65	70.82%
Kahlua Lopez	837	302	15.06	69.07%
Johnny A Seed	1253	343	14.58	68.56%
John Smith	897	191	11.71	64.26%
Anne Animus‡	635	277	15.39	51.07%
Mean	1193.17	370.00	16.06	70.56%
s.d.	280.67	127.63	3.80	4.46%

Note that the largest conversational model, Jackie Dauer, did not provide the most accurate responses. Nor did Vinny Devito, the conversational model in which the authors spent the most time. In fact, there is not a significant relationship between the accuracy of the conversational model and the total time spent on the model (Pearson Correlation $r^2 = 0.388$, $p = 0.186$); the breakdown of time spent per method and the relationship to accuracy is presented in section 4.4.4.3. The Marty Graw conversational model was the most accurate, and its accuracy is more than one standard deviation higher than the mean.

In the following sections, we reported an analysis of the *usage*, *quality*, and *efficiency* of stimuli sources, and evaluated how they relate to the overall accuracy of the conversational models. Throughout the results sections, we identified relationships, but did not attempt to clarify the implications of those patterns. We explain the implications of the patterns and lessons learned in section 4.4.6 Discussion.

4.4.4.1 Analysis of method usage

Percentage of conversational model from each method. The largest source of new stimuli was from the DKS method, followed by stimuli from the VHT method (Figure 4-7). Note that there is a large variance in the usage of each source (as indicated by the s.d. error bars in Figure 4-7), particularly for the Manual method. These numbers were calculated from the averages on a per model basis, which means each conversational model is weighted evenly, regardless of the size of the model.

The distribution shown in Figure 4-7 was not uniform throughout the created conversational models. The distribution on a per-model basis (Table 4-4) shows that DKS was the largest source of conversational knowledge for the Marty Graw, Jackie Dauer, Kahlua Lopez, and Johnny A Seed models, but not for the other three. The

Vinny Devito conversational model includes the largest percentage of stimuli from the manual method compared to the other VHs (Table 4-4). This shows that the Vinny Devito authors chose to use the Manual method more than the other groups. The John Smith model has a large percentage of stimuli from VHT, 10.04% more than the next highest. The maximum number of stimuli that can come from VHT is fixed. Since the number of stimuli is fixed, John Smith's high percentage of stimuli from the VHT method indicates that after using the virtual patient generator, John Smith's authors did not add as much to the conversational model as the other groups.

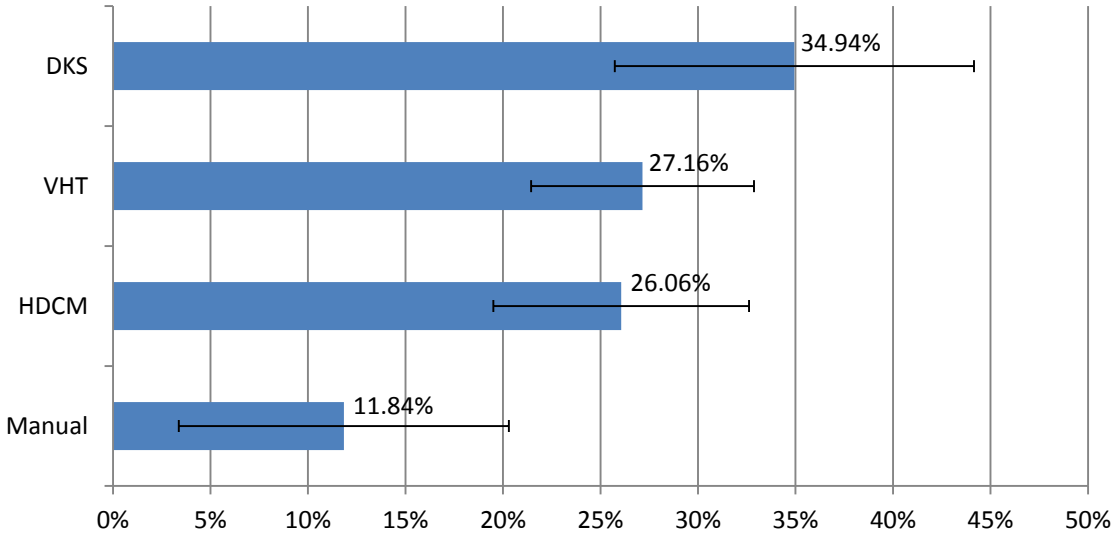

Figure 4-7. Mean percentage of stimuli in the conversational models from each method. DKS s.d. = 9.21%, VHT s.d. = 5.72%, HDCM s.d. = 6.56%, Manual s.d. = 8.46%, represented by the error bars

In contrast to the other six, the Anne Animus conversational model has a significantly larger percentage of stimuli from the HDCM method, and almost no stimuli from the DKS method (Table 4-4). This indicates that the authors of the Anne Animus conversational model could not make use of the DKS method. However, we note that a

large percentage of the conversational model still originated from VHT, indicating that they tried to use Conversational Knowledge Reuse.

Table 4-4. The percentage of stimuli from each source in each VH conversational model. ‡Shown for illustration purposes and not included in mean or s.d.

Name of virtual patient	Total Stimuli	Manual	HDCM	VHT	DKS
Marty Graw	1238	12.44%	28.27%	24.72%	34.57%
Vinny Devito	1358	25.26%	29.75%	24.23%	20.77%
Jackie Dauer	1576	7.92%	21.72%	22.74%	47.63%
Kahlua Lopez	837	2.15%	35.01%	28.20%	34.65%
Johnny A Seed	1253	17.48%	16.20%	24.82%	41.50%
John Smith	897	5.80%	25.42%	38.24%	30.55%
Anne Animus‡	635	3.31%	61.89%	32.44%	2.36%
Mean	1193.67	11.84%	26.06%	27.16%	34.94%
s.d.	281.49	8.46%	6.56%	5.72%	9.21%

Percentage of suggestions accepted from HDCM vs. DKS. We compare the usage of HDCM and DKS directly because both HDCM and DKS operate through the Suggestions System. However, the source of the knowledge from DKS may be less relevant than from HDCM and is expected to have a lower acceptance rate. For five of the six conversational models, HDCM suggestions were accepted approximately twice as often as DKS suggestions (as seen in Table 4-5).

The notable exception to this observation is the Jackie Dauer conversational model. The Jackie Dauer conversational model has a significantly higher percentage of accepted suggestions from both the HDCM and DKS sources (Table 4-5); this indicates that the Jackie Dauer authors were unusually likely to accept suggestions into the model. This behavior may have caused the Jackie Dauer conversational model to be the largest.

Table 4-5. The percentage of stimuli from HDCM and DKS that were accepted into each VH conversational model. ‡Shown for illustration purposes and not included in mean or s.d.

Name of virtual patient	HDCM Suggestions Processed	HDCM Suggestions Accepted	DKS Suggestions Processed	DKS Accepted
Marty Graw	429	76.92%	1150	36.78%
Vinny Devito	498	73.49%	812	34.61%
Jackie Dauer	348	92.24%	883	85.16%
Kahlua Lopez	444	64.41%	1014	28.50%
Johnny A Seed	264	74.24%	1001	52.85%
John Smith	287	75.61%	1136	24.38%
Anne Animus‡	475	78.53%	1617	0.99%
Mean	378.33	76.15%	999.33	43.71%
s.d.	93.31	9.04%	134.33	22.53%

Percentage of stimuli used from each method during interactions. While the VHT and DKS stimuli were the most represented in the conversational models (as shown in the previous section), they were not the most often used for responding during interactions (Table 4-6). The stimuli originating from the HDCM method were used approximately three times as often as the other methods. The stimuli from the other three methods were used a similar amount, with stimuli from the VHT method being used slightly more than the other two. There were also 9.17% of the utterances which received no response. To put the number of “no response” in perspective, with an average of 46.37 utterances per interaction, there was no response to an average of 4.25 utterances per interaction.

Note that the Vinny Devito conversational model has more hits against stimuli from the Manual method than the other models. Also note that the Anne Animus conversational model has significantly more responses from the HDCM source, zero responses from the DKS source, and the largest percentage of utterances with no response.

Table 4-6. Percentage of each stimulus source used for responses during interactions of each conversational model. ‡Shown for illustration purposes and not included in mean or s.d.

	Manual	HDCM	VHT	DKS	No Response
Marty Graw	14.29%	54.21%	12.45%	13.92%	5.13%
Vinny Devito	30.34%	38.06%	11.67%	10.77%	9.16%
Jackie Dauer	17.31%	40.64%	17.67%	17.67%	6.71%
Kahlua Lopez	3.44%	58.42%	17.18%	4.47%	16.49%
Johnny A Seed	14.03%	39.37%	21.72%	19.46%	5.43%
John Smith	4.96%	40.22%	20.94%	21.76%	12.12%
Anne Animus‡	15.84%	61.99%	9.50%	0.00%	28.51%
Mean	14.06%	45.15%	16.94%	14.67%	9.17%
s.d.	9.71%	8.79%	4.18%	6.36%	4.44%

Required overlap for Dynamic Knowledge Sharing usage. We have mentioned that the Anne Animus model is not similar enough to the other conversational models to make use of Dynamic Knowledge Sharing. To formalize the dissimilarity, we analyzed the overlap of the conversational models. We analyzed the overlap in two ways:

1) using the keyword vector matching algorithm used for simulations
2) using the overlap of keywords

Using the keyword vector matching algorithm used for simulations we took the stimuli from each conversational model, and checked if each other model had a response to that stimulus. If there was a response, it was marked as an overlapping piece of knowledge. Table 4-7 presents the percentage of overlapping stimuli in the conversational models. This data was computed bi-directionally by matching all each stimulus from each model against each of the other models -- Marty Graw may have a response to all of John Smith's stimuli, but the reverse may not be true. Excluding Anne Animus, the conversational models had an average of 85.01% overlap with s.d. =

7.59%. In contrast, the Anne Animus conversational model had an average overlap of 63.95% overlap with s.d. = 9.64%.

Table 4-7. Percentage of overlapping stimuli based on keyword vector analysis in the conversational models. Along the left is the source model, along the top is the model used for matching. ‡Shown for illustration purposes and only included in mean and s.d. specific to Anne Animus.

	Marty	Vinny	Jackie	Kahlua	Johnny	John	Anne	*Mean*	*s.d.*
Marty		89.7%	94.8%	82.7%	91.8%	81.9%	58.8%	88.2%	5.7%
Vinny	89.5%		93.1%	82.3%	90.9%	82.2%	61.7%	87.6%	5.1%
Jackie	76.0%	75.6%		72.6%	83.0%	68.0%	53.1%	75.0%	5.5%
Kahlua	85.4%	84.1%	96.1%		91.8%	80.4%	56.6%	87.6%	6.3%
Johnny	81.3%	81.5%	90.8%	75.7%		72.5%	50.1%	80.4%	7.0%
John	90.1%	90.9%	95.8%	87.1%	93.2%		63.0%	91.4%	3.3%
Anne‡	63.0%	67.1%	87.6%	67.9%	76.2%	62.4%		70.7%	9.6%
Mean	84.5%	84.3%	94.1%	80.0%	90.1%	77.0%	57.2%		
s.d.	5.9%	6.2%	2.2%	5.8%	4.1%	6.4%	5.0%		

We calculated the overlap of keywords by comparing the keywords in each of the conversational models. Words were designated as keywords if they were not stop words. Stop words are common words that convey little meaning to a sentence, examples include “the”, “which”, “that”, and “on”. The stop words list used in this computation is a custom list developed for healthcare scenarios. The overlap of keywords is presented in Table 4-8. Excluding Anne Animus, the conversational models had an average of 51.74% overlap with s.d. = 3.18%. In contrast, the Anne Animus conversational model had an average overlap of 34.36% with s.d. = 2.84%.

From this data, we see that the Anne Animus conversational model had less knowledge in common with the other conversational models than the others had with each other. This single data point provides only enough information to indicate that there is a threshold of information in common below which DKS is not useful. However,

this data does not provide enough information to indicate exactly where that threshold lays.

Table 4-8. Percentage of overlapping keywords in the conversational models. ‡Shown for illustration purposes and only included in mean and s.d. specific to Anne Animus.

	Marty	Vinny	Jackie	Kahlua	Johnny	John	Anne
Marty Graw		53.72%	50.05%	48.89%	55.12%	50.00%	30.65%
Vinny Devito	53.72%		51.58%	46.70%	55.28%	48.52%	31.82%
Jackie Dauer	50.05%	51.58%		50.05%	57.97%	47.16%	36.17%
Kahlua Lopez	48.89%	46.70%	50.05%		52.86%	52.86%	38.06%
Johnny A Seed	55.12%	55.28%	57.97%	52.86%		51.42%	33.52%
John Smith	50.00%	48.52%	47.16%	52.79%	51.42%		35.94%
Anne Animus‡	30.65%	31.82%	36.17%	38.06%	33.52%	35.94%	
Mean	51.56%	51.16%	51.36%	50.26%	54.53%	49.99%	34.36%
s.d.	2.70%	3.56%	4.03%	2.63%	2.51%	2.26%	2.84%

4.4.4.2 Analysis of method quality

Quality of stimuli from each method. To evaluate the *quality* of each method, we separated the overall conversational model accuracy into the accuracy per method. The DKS method had the highest average response accuracy, but also had a high standard deviation (Table 4-9). The VHT, HDCM, and Manual methods were all within 3% of each other. Given the overlapping standard deviations of the accuracies, we find that none of the methods can be said to be unequivocally higher quality than the others.

The knowledge from both VHT and DKS requires similarity of the scenarios, and the Anne Animus scenario was dissimilar from both the original scenarios from which the virtual patient template was generated and the other dysphagia scenarios. The Anne Animus conversational model had low accuracy for VHT. For VHT, only 33.33% of the responses were accurate, indicating the inappropriateness of the virtual patient template knowledge for the Anne Animus scenario. For example, the template contained a stimulus “how much do you weigh” and in interactions with Anne Animus,

interviewers were asking, “how much does your **baby** weigh”. Since the authors filled out the template regarding Anne, the responses were incorrect. As further evidence of the dissimilarity of the conversational model, the Anne Animus conversational model gave no responses using DKS stimuli, and so the DKS accuracy data is not applicable (N/A). This dissimilarity caused both a lack of knowledge reuse for DKS and low accuracy when the knowledge was attempted to be reused from VHT.

Table 4-9. Quality (accuracy) of each stimulus source for each conversational model. The “Model Accuracy” column indicates the overall conversational model accuracy. ‡Shown for illustration purposes and not included in mean or s.d.

	Model Accuracy	Manual	HDCM	VHT	DKS
Marty Graw	77.83%	89.74%	75.68%	85.29%	81.58%
Vinny Devito	72.84%	69.23%	71.70%	69.23%	96.67%
Jackie Dauer	70.82%	67.35%	75.65%	78.00%	76.00%
Kahlua Lopez	69.07%	100.00%	78.24%	70.00%	100.00%
Johnny A Seed	68.56%	67.74%	71.26%	77.08%	60.47%
John Smith	64.26%	50.00%	67.12%	76.32%	73.42%
Anne Animus‡	51.07%	71.43%	63.50%	33.33%	N/A
Mean	70.56%	74.01%	73.27%	75.99%	81.35%
s.d.	4.46%	17.92%	4.01%	5.89%	14.90%

4.4.4.3 Analysis of method efficiency

Efficiency of time spent on each conversational model. To calculate the efficiency of time spent on each conversational model, we divided the accuracy of the model by the time spent on that model to calculate the percentage gain per hour. The conversational model found to have the highest efficiency was John Smith (Table 4-10). John Smith also had the lowest accuracy. This finding may at first appear to discount this formula for computing efficiency; however, we note that the John Smith authors spent less total time than any of the other groups, which means a greater portion of their time was used on VHT. The second highest efficiency conversational model was Jackie Dauer. The Jackie Dauer conversational model also had the third highest accuracy.

Table 4-10. The efficiency of time spent on each conversational model. Efficiency is in accuracy gained per hour spent. Conversational modeling time and conversational accuracy are included for comparison. ‡Shown for illustration purposes and not included in mean or s.d.

	Modeling Time	Accuracy	Efficiency (per hour)
Marty Graw	17.49	77.83%	4.45%
Vinny Devito	22.84	72.84%	3.19%
Jackie Dauer	14.65	70.82%	4.83%
Kahlua Lopez	15.06	69.07%	4.59%
Johnny A Seed	14.58	68.56%	4.70%
John Smith	11.71	64.26%	5.49%
Anne Animus‡	15.39	51.07%	3.32%
Mean	16.06	70.56%	4.54%
s.d.	3.80	4.46%	0.75%

While we did not find a strong relationship between the total time spent and the accuracy of the conversational models, we did find a strong relationship between the time spent on HDCM and DKS and the accuracy of the conversational model (see Table 4-11 for method time and accuracy). HDCM to accuracy had a Pearson Correlation $r^2 = 0.793$, $p < 0.05$, and DKS to accuracy had a Pearson Correlation $r^2 = 0.859$, $p < 0.05$. Model accuracy, HDCM time, and DKS time were found to be normally distributed according to a Shapiro-Wilk test for normality. Note that Vinny Devito conversational model has the largest amount of time spent manually inputting conversational knowledge. Also note that the Marty Graw scenario had the largest amount of time spent on the DKS method.

Table 4-11. Time spent on each method for each conversational model. Author interactions are the author interacting with their own VH. Maintenance tasks include signing the license agreement, changing account settings, and changing character images. Conversational Model accuracy included for comparison. ‡Shown for illustration purposes and not included in mean or s.d.

	Manual	HDCM	DKS	VHT	Author Interactions	Maintenance	Accuracy
Marty Graw	2.82	2.50	6.70	0.28	3.56	1.62	77.83%
Vinny Devito	7.51	2.53	4.12	0.84	4.96	2.88	72.84%

Table 4-11 continued

	Manual	HDCM	DKS	VHT	Author Interactions	Maintenance	Accuracy
Jackie Dauer	3.62	1.57	3.98	0.23	3.02	2.24	70.82%
Kahlua Lopez	3.39	1.59	3.63	0.88	1.68	3.88	69.07%
Johnny A Seed	2.85	1.03	3.89	0.39	3.55	2.88	68.56%
John Smith	2.25	0.81	3.22	0.20	3.74	1.49	64.26%
Anne Animus‡	2.05	1.34	4.56	0.28	6.01	1.14	51.07%
Mean	3.74	1.67	4.26	0.47	3.42	2.50	70.56%
s.d.	1.91	0.72	1.24	0.31	1.07	0.90	4.46%

Efficiency of each method in number of stimuli per hour. The VHT method was found to be the most efficient method of adding stimuli to a conversational model (as shown in the *Mean* row of Table 4-12). Manual was found to be the least efficient method of adding stimuli to a conversational model. On average, VHT was approximately five times more efficient than HDCM, nine times more efficient than DKS, and twenty-four times more efficient than Manual.

Table 4-12. Efficiency of each method in terms of the number of stimuli per hour for each conversational model. ‡Shown for illustration purposes and not included in mean or s.d.

	Manual	HDCM	VHT	DKS
Marty Graw	54.65	139.94	1099.40	63.84
Vinny Devito	45.66	159.72	391.54	68.37
Jackie Dauer	34.52	218.61	1595.56	188.90
Kahlua Lopez	5.31	184.16	269.54	79.81
Johnny A Seed	76.89	197.84	800.86	133.66
John Smith	23.10	280.45	1712.62	85.15
Anne Animus‡	10.22	293.25	734.26	3.29
Mean	40.02	196.79	978.25	103.29
s.d.	25.01	49.48	602.16	48.80

Limits of stimuli per hour efficiency metric. The numbers in Table 4-12 reflect the raw efficiency of each method; but they do not take into account the limitations of the methods. For example, the virtual patient template used in the study has 320 stimuli.

Thus, while VHT may be able to provide a rate of 978.25 stimuli per hour, it provides a maximum of 320 stimuli total.

There is also a limit to the HDCM method based on the number of interviews conducted. HDCM can only draw knowledge from suggestions provided in the interviews. To illustrate this point, we use the Marty Graw conversational model. There were a total of 27 interviews with the Marty Graw VH in the three rounds of interactions. These interviews provided 429 suggestions through HDCM, 330 of which the authors accepted into the conversational model. With only the VHT, HDCM, and Manual methods, the Marty Graw conversational model would only have had 810 stimuli of its 1238 stimuli.

Part of the cause of this limited number of suggestions is the diminishing returns of interactions. If the suggestions are all processed, which makes the VH better at responding correctly to utterances, there will be fewer incorrect responses, and thus fewer suggestions gained for each additional interaction. The top four most accurate conversational models show diminishing returns for each round of development (Table 4-13), whereas the bottom two and Anne Animus show non-diminishing returns.

Table 4-13. Suggestions received per utterance for each conversational model. ‡Shown for illustration purposes and not included in mean or s.d.

	Round 1	Round 2	Round 3
Marty Graw	0.441	0.176	0.116
Vinny Devito	0.306	0.146	0.144
Jackie Dauer	0.283	0.221	0.164
Kahlua Lopez	0.388	0.255	0.205
Johnny A Seed	0.230	0.119	0.184
John Smith	0.374	0.190	0.247
Anne Animus	0.282	0.366	0.333
Mean	0.337	0.184	0.177
s.d.	0.078	0.049	0.046

Examining the suggestions received per utterance of the Vinny Devito model, we see an additional negative effect of using the Manual method early in the modeling process. After Round 1, the suggestions received per utterance for the Vinny Devito model dropped significantly. This drop may be due to the time the authors spent manually entering stimuli. That manual work may have diminished the returns of the HDCM and DKS methods (both more efficient methods), thus lowering the overall efficiency of their modeling process, and increasing the time required.

The limited resource of novice interactions and HDCM's diminishing returns are the reason for the DKS method. Continuing our example of the Marty Graw model, the DKS method provided an additional 1150 suggestions, 423 of which the authors accepted into the conversational model. While the DKS method is subject to the same diminishing returns as the HDCM method, the DKS method extracts additional knowledge from every interview, thereby enhancing the overall knowledge acquired from every minute of novice time. And while there were only 27 interactions with the Marty Graw VH, DKS simulated the Marty Graw VH with the 159 interactions of the other VHs. Using all four methods in combination allowed the Marty Graw conversational model to reach 1238 stimuli and 77.83% accuracy within the limited number of novice interactions.

4.4.6 Discussion

The results of this evaluation indicate that students within a course can now generate VHs using HDCM with Conversational Knowledge Reuse. All six conversational models (excluding Anne Animus) were as dense as the corpuses of virtual patients created by domain experts and used in classroom instruction (Rizzo et

al., 2010; Rossen and Lok, 2012). That is, the six groups created VHs of complexity potentially usable for classroom instruction.

The total time required by these groups of students to create a VH was similar to the time spent by a single expert previously (11-25 hours, Chapter 3) (Rossen and Lok, 2012). However, taking into account the cost of each student learning the Virtual People Factory system and their unfamiliarity with the medical content, a similar amount of time spent by a group of students may indicate a faster overall time compared to a single expert. These findings indicate that the use of VHT and DKS decreased the total time to model the conversations.

4.4.6.1 How to most effectively spend time when conversational modeling

Given the high efficiency of time spent on VHT (Table 4-14), we find that VHT was the most effective method of acquiring conversational knowledge.

Table 4-14. Summary of the *usage*, *quality*, and *efficiency* of each method.

	Usage (model)	Usage (responses)	Quality (accuracy)	Efficiency (stimuli)
VHT	27.16%	16.25%	75.99%	16.3 per minute
HDCM	26.06%	44.16%	73.27%	3.3 per minute
DKS	34.94%	14.24%	81.35%	1.7 per minute
Manual	11.84%	15.90%	74.01%	0.7 per minute

The next most effective method was HDCM because of the high efficiency of the method as well as the high usage of the knowledge during interactions – see "Usage (responses)" in Table 4-14. After VHT and HDCM, it is not as obvious which is more effective between Manual and DKS. While the Manual method has a slightly higher usage during interactions, the DKS method has a higher quality and efficiency. The deciding factor was the relationship between time spent on DKS and the overall quality of the conversational model. This relationship indicates that DKS is the more effective

method. Table 4-14 presents a summary of *usage*, *quality*, and *efficiency* of each method, ordered by where time was most effectively spent.

4.4.6.2 The effect of the Conversational Knowledge Reuse methods

From the study results we see that the Conversational Knowledge Reuse methods (VHT and DKS) have improved the efficiency of conversational modeling and thereby facilitated a greater coverage of the conversational space in a shorter period of time. The effect becomes apparent when we examine the example of Anne Animus. Because the Anne Animus scenario was different than the other dysphagia scenarios, both the VHT and DKS knowledge was ineffective. The effect was that the Anne Animus authors were using only the HDCM and Manual creation methods to generate the conversational model. Unfortunately, with only these tools, the resulting Anne Animus conversational model was of relatively low quality (accuracy = 51.07%). This is despite the authors using a similar amount of time compared to the other groups (15.39 hours vs. mean of 16.06 hours). We acknowledge that the time taken by the Anne Animus group to examine and discard suggestions from the DKS method (4.56 hours) contributed to the total time used on conversational modeling. However, the Anne Animus group did process all of the HDCM suggestions they received and spent 2.05 hours manually adding stimuli and responses, yet Anne still had no response to 28.51% of the utterances. The fact that they processed all of their HDCM suggestions indicates that without Conversational Knowledge Reuse there was not enough information available to enumerate the conversational space and create a robust conversational model.

4.4.6.3 Comparison of the six conversational models

The following is a discussion comparing and contrasting each of the conversational models based on their *usage, quality,* and *efficiency*. The highest quality conversational model in the study was Marty Graw (accuracy = 77.83%). The Marty Graw authors made efficient use of their time by putting over 50% into processing suggestions (HDCM and DKS), the highest percentage of all the groups. This distribution of time contrasts with the time spent by the authors of the Vinny Devito conversational model. The Vinny Devito authors achieved the second highest quality (accuracy = 72.84%), but the model required over 20% longer than the Marty Graw conversational model (22.84 hours vs. 17.49 hours), and achieved 5% less accuracy. The additional time spent by the Vinny Devito authors was caused by putting 33% of their time (7.51 hours) into the Manual method, whereas the Marty Graw authors spent the least time on the Manual method (2.82 hours, 12%). The Vinny Devito authors put Manual time into the model early in the conversational modeling process, and thereby lowered the effectiveness of HDCM and DKS by causing fewer suggestions to be gathered. The contrasting examples of Marty Graw and Vinny Devito indicate that while one can make progress using the Manual method, it is more efficient to spend time on the HDCM and Conversational Knowledge Reuse methods.

The third most accurate conversational model was that of Jackie Dauer. The Jackie Dauer authors used less time (14.65) than the Vinny Devito authors to achieve a similar (within 2.2%) accuracy (70.82%). Jackie Dauer's authors achieved this accuracy by accepting a high percentage of the suggestions from both HDCM (92.24%) and DKS (85.16%) compared to the average (HDCM = 76.15% and DKS = 43.71%). By accepting suggestions that were off-topic, and giving their VH a response to those

suggestions, their conversational model received a boost in efficiency (second highest efficiency with 4.83% accuracy per hour) which resulted in an overall high accuracy for a relatively low time spent. In contrast, for the next highest accuracy model, Kahlua Lopez, the authors took the opposite approach than the Jackie Dauer authors. They accepted a low percentage of HDCM suggestions (64.41%) and a low percentage of DKS suggestions (28.50%). They also added very few manual stimuli and responses (3.44%). Yet they achieved only a slightly lower accuracy (69.07%) than the Jackie Dauer authors did in approximately the same amount of time (15.06 hours). This data indicates that the Kahlua Lopez authors were very careful with the stimuli they added to the conversational model. The effect was that Kahlua Lopez had no response to utterances more often than any other VH (16.49%), but when she did respond, the responses were highly accurate. The two strategies represented by Jackie Dauer and Kahlua Lopez may be appropriate at different stages in the conversational modeling process.

The Johnny A Seed conversational model does not stand out in any measure. Accuracy was near the middle (68.56%) and modeling time was similar to Jackie Dauer and Kahlua Lopez (14.58). It may be that the consistent middle ground strategy is a less successful than either accepting a high percentage of the suggestions or being careful with the suggestions accepted.

The lowest accuracy conversational model was John Smith. The John Smith authors spent fewer hours on modeling the conversation than any other group (11.71 hours). However, the John Smith model had the highest efficiency of accuracy gained for time spent (5.49% per hour). John Smith's high efficiency can be attributed to VHT.

Since the John Smith authors spent less time processing suggestions or manually editing than the other author groups, they received a greater proportion of their efficiency from VHT.

Overall, we see that different methods are useful at different stages of modeling the conversation. In the first stage, VHT is the most effective, and rapidly seeds the HDCM with DKS process (evidenced by the modeling efficiency of John Smith). In the beginning of the HDCM with DKS process, it may be effective to accept only the suggested stimuli that are highly pertinent to the conversational model in order to keep the number of suggested stimuli coming from each interaction high (evidenced by the accuracy of Kahlua Lopez). In the later stages, it may be most effective to accept nearly all of the stimuli into the conversational model to achieve maximum coverage of the conversational space (as evidenced by the accuracy of Jackie Dauer). The Manual method is the least efficient, and should be avoided in the early stages because of the time and efficiency costs, but may be helpful as needed in the later stages to achieve a higher overall coverage of the conversational space (as evidenced by the total time and accuracy of Vinny Devito). In conclusion, we find that the conversational modeling methods are most effective when used in order of their efficiency, first VHT to seed the model, then HDCM with DKS to refine the model, and then Manual to improve the content of the conversational model.

4.5 Technical difficulty of the methods vs. their efficiency

This section compares the technical difficulty in implementing each of the conversational methods to the efficiency gained by the VH author. The relative ease of implementation and its relationship with efficiency may inform computer engineers

where to spend engineering resources. The difficulty of implementing the modeling methods appears to be proportional to their efficiency.

The Manual method is the easiest to implement, but only allows Centralized Conversational Modeling, and is the least efficient method of modeling a conversation. The DKS method is easy to implement, but only after HDCM is implemented. DKS only requires the background simulation of previously conducted interactions.

The HDCM method requires two complex systems, the tracked interaction system and the suggestions system. The interaction system records new stimuli and facilitates users reporting incorrect responses. The interaction system also feeds new knowledge into the suggestions system where it is processed by VH authors. The suggestions system itself has several features that facilitate efficient processing of conversational knowledge (see section 3.2.6).

The most efficient method, VHT, also requires the most implementation and requires knowledge engineering of the templates. The implementation of an efficient VHT system requires a complex database backend, an interface to customize the templates, and the generation of those templates into a conversational model. Despite the cost of engineering, the VHT method is still effective overall, because computer engineer and knowledge engineer time is leveraged to improve the efficiency of a large number of VH authors. It is also important to note the dependency of the conversation knowledge reuse methods on HDCM and Centralized Conversation Modeling to generate the original knowledge that can be reused. Conversational knowledge reuse methods, especially VHT, work only when there is conversation knowledge that has been already acquired. For example, if you would like to apply the conversational

modeling methods discussed in this paper to a new domain, like training school teachers to have better interactions with children, you will have to build the first few conversational models using HDCM or Centralized Conversational Modeling. Only when sufficient knowledge has been acquired through these methods can a template be created for VHT implementation.

4.6 Current Data and Future Work on Factors Influencing the Speed of Conversational Modeling

This section describes the variables that determine the speed of conversational modeling using HDCM with Conversational Knowledge Reuse. When modeling conversations for question-answering VHs the complexity of the scenario and the availability of resources are central to the efficiency of the conversational modeling methods. We discuss what is currently known about the limitations of these resources and indicate areas where future research will be conducted to further understand the effect of these variables on conversational modeling.

4.6.1 The Complexity of the Scenario

In healthcare, interviews have varying levels of complexity. The complexity is determined by the health problems the virtual patient has, the number of medications the patient is taking, the type of health problems experienced by the patient's parents, and the other topics covered by the scenario. The added complexity for each additional topic is influenced by the complexity of the new topic as well as its interactions with the other topics. These interactions cause the required corpus to grow non-linearly with the number of topics.

Differences in the complexity of new topics are evident if we examine adding a new medication to a patient scenario. The number of additional questions asked for

each additional medication is dependent on the issues caused by the medication. In the Vic Johnson scenario of Chapter 3, Vic was taking four medications: Zestril, Synthroid, Tums, and Aspirin. The number of stimuli asked about each of these medications varies (as seen in Table 4-15). Taking too much Aspirin caused Vic's chief complaint (an Ulcer), which required more questions to be asked about Aspirin as compared to the other medications. The symptom of stomach pain from Aspirin is simpler than some other medication side effects. Had Vic been taking too much Zestril instead of too much aspirin, he would have had low blood pressure and kidney failure, which would have caused a wider set of secondary symptoms, and likely even more questions about Zestril than were asked about Aspirin. The number of additional stimuli required for one additional medication is determined by the implications of that medication on the scenario.

Table 4-15. Number of stimuli for medications taken by Vic Johnson

	Stimuli
Aspirin	252
Zestril	177
Synthroid	149
Tums	118

The added complexity of a new topic is also dependent on interactions with other topics of the scenario. For example, if Vic Johnson were also taking Ibuprofen, another medication that may cause stomach pain, there would have been many questions about Ibuprofen; but it would also increase the number of questions about Aspirin. These additional questions about Aspirin would be necessary to distinguish the side effects of Aspirin from the side effects of Ibuprofen.

As a general heuristic, these examples indicate that interactions between the topics cause the complexity of the scenario to grow non-linearly with the number of topics. Examining this as a mathematical system, if we consider the selected topics of the scenario to be the independent input variables and the number of required stimuli in the corpus to achieve accuracy to be the dependent output variable, we see that the required number of stimuli (output) is not directly proportional to the number of topics (inputs). The addition of one new topic may require new stimuli on many other topics to distinguish the new topics from the existing topics, and would thereby increase the scenario's complexity in a non-linear fashion. This growth pattern warrants further research into the limits of the current technology to represent complex interactions.

4.6.2 The Availability and Motivation of Novice Interviewers and VH Authors

The conversational modeling methods require interactions with end-users (generally novices) to develop the conversational models. The speed of development depends in part on the number of novice users and their motivation to conduct the interaction. The more novice users, and the longer the conversations, the more knowledge will be gathered. However, there are some considerations with how both novice interviewer time and VH author time can be used most effectively.

Novices are unlikely to be willing to interact with the same VH multiple times and there are a limited number of novice interviewers available. We should therefore strive to acquire maximum knowledge from each novice interaction. In order to get more depth in the conversation and maximize knowledge gathering, VH authors should process all of the knowledge as it arrives during interactions. Processing knowledge immediately causes future questions to be answered correctly and triggers the interviewers to ask

follow on questions, thereby gathering additional knowledge. Though processing knowledge immediately appears ideal, it is also unrealistic.

To make knowledge processing more efficient for the VH author, VH authors generally gather a set of suggestions, and process the set all at once. Processing sets of suggestions at once saves the VH author time by allowing them to get into a flow of processing suggestions. To determine the correct number of interactions in each of these sets, developers will weigh the cost of novice interactions against the cost of VH author time. Making this calculation will require studying the amount of time saved by processing batches of suggestions at once as well as the amount of knowledge lost by waiting to process novice suggestions.

In the Conversational Knowledge Reuse study, we required the VH authors to separate their development into three rounds. This separation lowered the difficulty of analyzing and presenting the study results, but may not be the most efficient method of gathering novice knowledge. In practice, VH authors may achieve better results by processing the knowledge as soon as the knowledge is available and they have time to process it. Future research will examine the differences in efficiency between processing large blocks of suggestions and processing suggestions more frequently.

4.6.3 Improvement from Novice Interactions

We examine the effect of novice motivation and VH author availability using an example from the Dysphagia Management study (section 4.4). The Marty Graw conversational model achieved the highest accuracy of the seven conversational models in the study. An accuracy of 77.83% was achieved using 27 interactions with Marty Graw as well as 159 simulated interactions using DKS. In interactions directly with Marty Graw, Marty had improved accuracy of responses with each round of testing

and development. This improvement provided a diminishing number of new stimuli per interaction for each subsequent round of testing (Table 4-16). The rounds had 8-10 interactions plus 53 simulated interactions per round. Given these interactions the Marty Graw scenario gained 16-20% accuracy per round of development. These gains in accuracy depended on three factors: complexity of the scenario, motivation of the novices, and availability of the VH authors.

Table 4-16. Number of stimuli and accuracy of responses during interactions with Marty Graw for each round of testing

	N	Accuracy of responses	Average number of questions	Average number of new stimuli
Round 1	8	42.17%	55.00	24.25
Round 2	10	58.47%	59.20	10.40
Round 3	9	77.83%	41.11	4.78

The complexity of the Marty Graw scenario is indicated by the topics covered in the conversation (Table 4-17). The complexity of this conversation lies in the required depth in each topic as well as the interaction between topics. The scenario contained both deep and shallow topics as well as topics that interact. The Chief Complaint contains deep topics covering the description of the issue (swallowing pain), the onset of the issue, and the social implications of the issue. The family history contained shallow topics because the father's heart attack and the mother's diabetes do not pertain to the chief complaint, and so do not warrant further exploration by the interviewer. The medications and social history are examples of topics that interact. The medication history interacts with the chief complaint, and so was explored in depth because each medication might cause a swallowing disorder if taken improperly. The social history also interacts with the chief complaint because eating is a significant part of Marty Graw's career as a Gumbo Chef.

Table 4-17. Topics covered in the Marty Graw conversation

Subject	Topics of conversation
Chief Complaint	Swallowing pain caused by esophageal stricture, diet changes to accommodate pain
Medication History	Avapro, Nexium, Tums, and Motrin
Medical History	GERD, high blood pressure, tonsillectomy at 18 years, allergic to tree nuts and penicillin, 55 years old
Family History	Father died of a heart attack, mother has diabetes and high cholesterol
Social History	Gumbo Chef, married with two daughters, BA in culinary arts, stress from financial problems, stress from work, stress from family
Risk Factors	Drinks wine daily, previously smoked but does not now, overweight, weight loss caused by diet changes, no exercise, no illegal drugs
Greeting/Exit	Introduction and exit

The second factor is the motivation of the novice interviewers. The novices involved in the development of the Marty Graw had higher motivation than in previous VH developments. Marty Graw's interviewers were externally motivated because the interviews were a part of their coursework, and were internally motivated because the interviewers were also VH authors themselves. The interviewers asked an average of 51.93 questions per interaction. As a comparison, in the development of Vic Johnson (Chapter 4), the novices asked an average of 37.95 questions per interaction. Internal and external motivation strategies have been previously explored for developing conversational models and can be used to increase the motivation of novice participants (Halan et al., 2010).

The last factor is the availability of the VH authors. Because this development was done during a study, the VH authors were "available" only during the development phases of the process, that is, between rounds of interactions (see section 4.4.3 Procedure). If the authors were available more frequently (such as after every interaction), we may have observed larger gains for each novice interaction. We base

this expectation on the understanding of how these conversational models grow; with earlier processing of novice knowledge, the interviewers would have been able to explore more depth in the conversation, and therefore provided greater depth of knowledge acquisition.

Together the complexity of the conversation, the motivation of the novices, and the availability of the VH authors determined the speed of accuracy improvement when developing the Marty Graw conversational model. The conversation is relatively complex, and is typical of a VH patient scenario; the motivation of the novices was higher than in the previous study developing Vic Johnson (Chapter 4); and the participation of the authors was similar to the previous study. To provide a greater understanding of the effect of the complexity of the conversation, the motivation of novices, and the availability of VH authors, future research will examine varying levels of conversational complexity, strategies for participant motivation, and differing patterns of author participation.

4.7 Potential for New Application Areas

The Conversational Knowledge Reuse methods described in this chapter may open new application areas. Because these VHs can be generated more quickly, they may be useful for applications where rapid generation is essential. For example, with rapid conversational modeling methods, VHs could be used for emergent situations such as an epidemic of infectious disease. VH authors could rapidly generate a set of VHs to represent patients displaying signs of the infectious disease, and healthcare practitioners could use these VHs to learn to recognize the symptoms. This system may also be beneficial for using the VH creation process as an educational tool. For example, during the study presented, we also examined if the students improved their

ability to understand patients. We found that students did improve their ability to perceive the personality of a conversational partner (Halan et al., 2012).

CHAPTER 5
CONVERSATIONAL KNOWLEDGE REUSE: USING VIRTUAL HUMANS TO BOOTSTRAP THE CREATION OF OTHER VIRTUAL HUMANS

This chapter presents the analysis, design, implementation, and evaluation of a novel method for using virtual human knowledge to bootstrap the creation of new virtual humans, *Virtual Human Bootstrapping* (research questions 1-3). It also demonstrates that this Conversational Knowledge Reuse technique opens a new application of virtual human interpersonal simulation. The implementation of *Virtual Human Bootstrapping* method is called The Roleplay Trainer Creator. The Roleplay Trainer Creator uses existing virtual human conversational knowledge to generate new virtual humans. We describe a study which uses The Roleplay Trainer Creator to generate virtual human medical students. These virtual human medical students are then used for training standardized patient actors to conduct practice medical interviews. The design and evaluation of *Virtual Human Bootstrapping* was published in the proceedings of the International Conference on Intelligent Virtual Agents (Rossen et al., 2010).

Personal Contributions. I conceptualized the method, implemented the Roleplay Trainer Creator, and designed and analyzed the study.

Collaborators. Dr. Juan Cendan was involved in the discussions of the bootstrapping method and provided valuable reviews and feedback during the development of The Roleplay Trainer Creator. Dr. Cendan also provided access to study participants (standardized patients) and served as the domain expert author in creating the virtual human medical student used in the study.

Relevance to thesis. This chapter describes a method and system that demonstrates an additional application of Conversational Knowledge Reuse to rapidly

generate virtual human conversational models, and opens up a new application area for virtual humans.

5.1 Overview

Across a wide range of fields, interactions with conversational virtual humans (VHs) are becoming a part of training interpersonal skills. Medical students practice taking a medical history by conversing with a VH patient (Johnsen et al., 2005), teachers re-enact classroom situations with virtual students (Dieker et al., 2007), and soldiers learn conflict resolution in interactions with virtual civilians (Hill et al., 2003). As described in section 1.3.3, during development of these human-VH interactions, the role played by the human and the role played by the VH are fixed. If these roles could be reversed (e.g. a human patient with a VH doctor) it would facilitate interpersonal skills training for additional large populations (Rossen et al., 2010). Traditional methods of creating a conversational model for the reversed role would double development time and effort. To solve this problem, we propose a new Conversational Knowledge Reuse method, *Virtual Human Bootstrapping*, which allows authors to rapidly reverse the roles in these conversations, thereby allowing VHs to play the previously human side of the interaction.

The *Virtual Human Bootstrapping* method uses recordings of human-VH interactions to bootstrap the creation of a new VH representing the human. When a human-VH interaction is conducted, recordings are made of the human user's utterances. These recordings ("interaction logs") are then used to generate new VHs to take the (originally) human role in the conversation. This is accomplished by aggregating the scenario's interaction logs. Given a set of interaction logs having

sufficient size and scope, we can create a conversational model for the human role in the human-VH interaction.

The *Virtual Human Bootstrapping* method was implemented in a web-based application called The Roleplay Trainer Creator. We conducted a study investigating use of The Roleplay Trainer Creator to apply this bootstrapping technique to medical education. In medical education interview training, standardized patients (human actors) are paid to roleplay a patient for medical student education. To help train standardized patients for standardized patient interactions, we propose to create virtual medical students from hundreds of interactions of medical students with a virtual patient. The virtual medical student generation process is as follows (see Figure 5-1). First, medical students interact with a virtual patient, resulting in a set of interaction logs. Next, those logs are used by The Roleplay Trainer Creator to generate a VH medical student. Last, that VH medical student is used to train standardized patients. We report on a pilot study using a generated virtual medical student to train standardized patients.

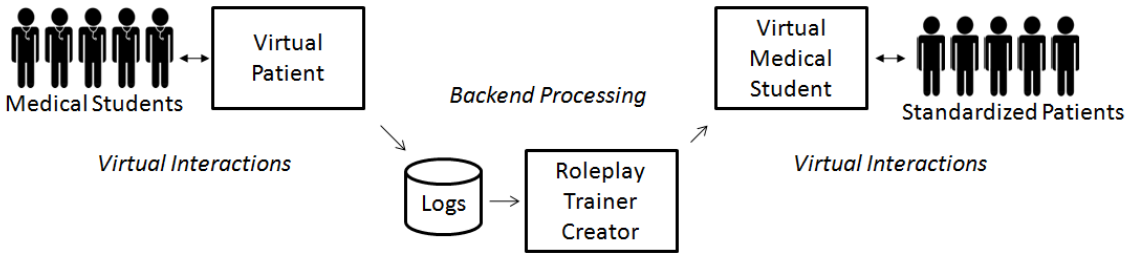

Figure 5-1. The virtual human medical student creation process, an example of virtual human bootstrapping

Driving application: healthcare standardized patient training. Standardized patient interactions have been used since 1963 for training medical interview skills. The medical interview is a 5 to 30 minute conversation in which the healthcare professional and patient exchange information. The healthcare professional elicits both verbal and

physical symptoms from the patient -- for the purposes of this research we focus only on the verbal content. Standardized patients are hired actors that roleplay a medical condition. They interact with many medical students (i.e. hundreds of students over several days, one student at a time) and need to represent the same patient in each student interaction. Representing the same patient with each student is crucial to providing standardized (i.e. *equal*) education and evaluation of students.

Standardizing standardized patients is a difficult problem (Lei et al., 2004; Tamblyn et al., 2009; Walters et al., 2005; Wessel et al., 2003). On average, standardized patients respond with 90.2% correct verbal content of responses, but their accuracy can be as low as 30% (Tamblyn et al., 2009). Reasons for these accuracy errors include: forgetting the correct answer, misunderstanding the question, or use of preparation materials that do not provide an answer to the student's question.

Typically, standardized patients are trained by personnel that are experienced in the field of healthcare simulation. At the University of Florida, healthcare simulation experts spend one hour reviewing written materials that pertain to the case with the standardized patient. The written materials are approximately ten typed pages. The case material contains instructions on dress, conduct of emotion, and a description of the health history (e.g. Smoking: 2 packs a day). Standardized patients are also shown videos of interactions between medical students and other standardized patients conducting the medical interview. In most instances, the standardized patient then takes that written material home and commits the details to memory in an effort to portray the character consistently.

The standardized patients grasp the specific goals of the scenario but can exhibit performance variability, particularly when specific instructions have not been provided (Tamblyn et al., 2009). For example, a student may ask “do you drink alcohol?” and if a particular answer to this question has not been presented by the case materials, the standardized patient has the ability to answer that as they see fit. In many cases this may not be a problem; however, a response of “I drink every day” could persuade the interviewing student that alcohol plays a particular role in the scenario when, in fact, it may be irrelevant. These types of details are very important to the standardization of training and testing. The current methods for training leave out some of these specifics due to the large burden of information as it is currently presented, and the passive nature of engagement of the standardized patient as they are largely reading these materials with varying levels of supervision.

Preparation materials given to standardized patients are also sometimes incomplete because the material creators are unable to anticipate all questions that medical students may ask. The reason for this is that the training materials are generated by medical interviewing experts. We have found that experts cannot predict the many paths that novices will pursue in a medical interview. In our experience with developing VH patients from 2005 to 2011 (Johnsen et al., 2005), the contents of patient cases are only the essential information the patient is supposed to convey rather than an exhaustive set of what will be asked.

This limited set of data results in standardized patients:

- Making up responses to questions they don’t know, and
- Changing their answers between interactions because they do not know what the "correct" answer is.

With additional training using a virtual medical student, we may be able to increase the standardization of standardized patients. We propose a method for generating virtual medical students derived from the questions real medical students asked virtual patients. As virtual patients are modeled after standardized patients, we anticipate that the questions students ask virtual patients will be predictive of the questions students will ask standardized patients. Given a dataset that covers the space of questions medical students will ask, we can generate a representative virtual medical student. The virtual medical student's questions will be those most commonly used in the dataset as well as the critical but unusual questions. Since the questions came from medical students themselves, they will be phrased in the manner medical students would use. Thus, standardized patients may be able to interact with the virtual medical student to prepare for interactions with real medical students.

The Roleplay Trainer. The proposed standardized patient trainer, the Roleplay Trainer, allows for engagement of the standardized patient with a system that mimics a dialogue and forces the discovery of previously defined critical-information during that dialogue (details in section 5.2). The Roleplay Trainer provides the standardized patients with:

1. More complete coverage of the questions that medical students will ask,
2. The correct responses to those questions,
3. Experience answering the questions, and
4. Feedback on their responses to those questions.

The Roleplay Trainer could serve as a method to enhance recall of important details and minimize the variability in responses that are presented by the standardized patient. By conversing with the virtual medical student, standardized patients may be better able to standardize their interactions with real medical students, providing

increased educational value to the students and more accurate evaluation of medical student competency.

5.2 Roleplay Trainer Creator: Generating Virtual Versions of the Human Partner

5.2.1 Overview

The Roleplay Trainer Creator enables domain experts to create question-asking VHs from the interaction logs of question-answering VHs. These question-asking VHs are used to train roleplay partners (standardized patients) to conduct a question-answering conversation (such as a medical interview). The challenge with generating a question-asking VH that accurately represents a specific class of human (a medical student) within a specific type of conversation (a medical interview) is determining:

1. The representative questions to ask. The VH needs to ask not only the common and to the point questions, but also uncommon and possibly off topic questions.
2. The order in which to ask them. The VH asks questions in an order that is appropriate for a medical student, rather than a medical expert.

The Roleplay Trainer Creator enables a domain expert to create a VH roleplay partner by assisting the expert in selecting representative questions, and ordering those questions appropriately. The system extracts questions that were actually asked by users by analyzing the interaction logs of question-answering VHs.

The original question-answering VHs were developed using a collaboration between end-users, and domain experts using Human-centered Distributed Conversational Modeling. These VHs went through conversational model development until they had a robust ability to answer questions. We can infer that because these VHs have a robust ability to answer questions, the logs of these conversations enumerate the space of both sides of the conversation. Using these log files, we can extract the questions for the question-asking side of the conversation.

There are two challenges to generating the question-asking side of the conversation from the question-answering log files:

1. **Handling a large dataset:** The log files from developing question-answering VHs are too large for a human to parse by hand. The test-bed VH used in this study has 694 interactions of medical students interacting with a VH patient. These interactions consist of 14872 questions asked by medical students. Many of these questions are semantically identical to each other. The semantically unique questions must be identified and their usage analyzed as part of reducing those 14872 questions into a size appropriate to a single conversation.
2. **Ordering the questions:** The questions should occur in a plausible order. The order should be similar to the order used by representatives of the intended group. For our example, the virtual medical student should ask questions in the same order a real medical student would ask them.

These two challenges are addressed in The Roleplay Trainer Creator (detailed in section 5.3.1). The Roleplay Trainer Creator generates a script that is stored in a database. The Roleplay Trainer script is then used by Virtual People Factory to simulate the question-asking VH in a browser-based interaction. Details on The Roleplay Trainer Creator and simulating the VH are given below.

5.2.2 System Implementation

The Roleplay Trainer Creator is a web-application developed using PHP, JavaScript with jQuery, MySQL, and Apache. It was developed on top of Virtual People Factory web development framework (described in section 3.2) (Rossen et al., 2009). The Roleplay Trainer Creator generates conversational models that are compatible with the Virtual People Factory scripting system.

5.2.2.1 Selecting questions

The Roleplay Trainer Creator enables question selection by pre-processing the interaction logs and then enlisting the expert to validate the results. This is a human-centered approach in that it leverages computational processes (filtering and ordering)

to do the tedious work, and then leverages human-intelligence for only the portions of the process that need human intervention.

There are two categories of questions to select, the most common questions and the critical questions. Selecting the common questions is accomplished by:

1. Identifying the semantically identical questions,
2. Analyzing the usage of the questions in the interaction logs, and
3. Presenting the questions and responses to a domain expert for final selection

To identify semantically identical questions, each question asked was categorized by its similarity to questions that were previously validated for use in the existing question-answering VH conversation. Similarity was determined using lexical keyword matching. Matching is performed by comparing the low frequency words, or keywords, in a question to the low frequency words in the validated questions in the VH conversational model.

Next, each question in the transcript logs is compared to previously validated questions in the question-answering conversational model. The question in the question-answering VH conversational model that has the highest similarity to the question being examined is categorized as semantically identical. Using the test-bed VH conversational model it was previously found that the accuracy of this algorithm achieves 80% for new material and 90% for training data (Rossen et al., 2009). The semantically identical questions were then counted once for each log file they appeared in and divided by the total number of log files to determine the frequency of usage.

The frequency of usage is displayed in the interface for selecting which questions to use in the question-asking VH seen in Figure 5-2. Experts place check marks next to the questions to select them. They can also choose to select all of the questions used in

more than X% of the interactions. With this interface, the most common questions can be quickly selected. In essence, this process rapidly generates a “most common questions” medical student. The top eight most common questions from the test-bed scenario can be seen in Figure 5-2.

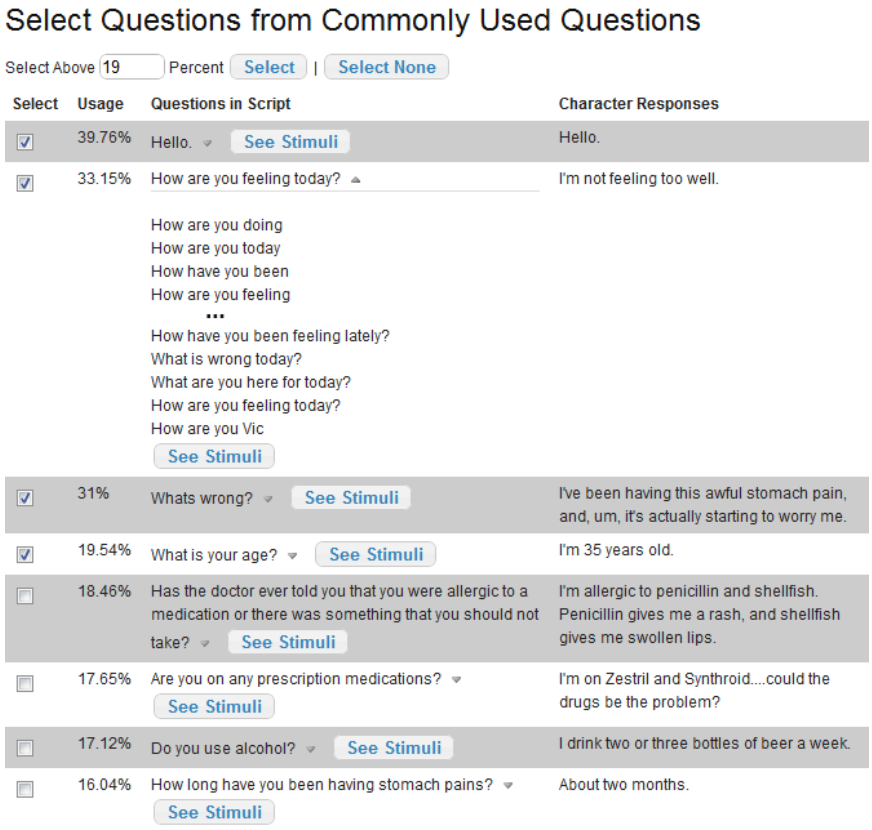

Figure 5-2. The Roleplay Trainer Creator interface for selecting questions by usage

The example shows the creation of the virtual medical student used in the pilot study. “Questions in Script” are the questions from the question-answering script to be selected for the question-asking script. “Character Responses” are the original responses from the question-answering script. Below the question “How are you feeling today?” is a list of expert validated semantically equivalent questions. The uncommon, but important, questions are needed as well. This is where the expert is essential; they can look through the rest of the questions and select the important questions.

Selecting questions for the test-bed scenario. For the question-answering VH used in our study, there were 493 semantically unique questions. The appropriate number of questions to use for the question-asking VH was determined by examining the average length of a typical virtual patient interview in the test-bed scenario. After eliminating outliers, the average length was 40 questions ± 27. Using the upper end of the average length plus standard deviation, we determined that approximately 70 questions was an appropriate length to maximize the number of questions the system could ask the standardized patient while remaining a realistic length. The expert was able to select 71 questions for our virtual medical student. This number was a compromise between maximizing coverage, minimizing expert time to select questions, and having a reasonable interaction length. The expert first selected the top 19 most commonly used questions (questions used in > 8.5% of interactions), and then selected individual questions from the rest of the list. For example, in our test-bed scenario the following questions about aspirin were important, but were asked by less than 2% of the users:

- “Do you take aspirin with food?”
- “Did the pain start after taking the Aspirin?”
- “Have you spit up blood?”
- “What side-effects are you experiencing from Aspirin?”

Selecting the frequently used and important questions took the expert approximately 60 minutes. Next, the questions were ordered.

5.2.2.2 Determining question order

The next challenge was determining the question ordering. The system again pre-processes the data, and enlists the expert to validate the results. The goal is to have the VH roleplay partner (virtual medical student) ask questions in an order similar to the

order used by real novices (medical students). Real novice users may take a meandering route; this is in contrast to how an expert would conduct the interview. Expert interviewers tend to choose an optimal path to arriving at a diagnosis, and so skip extraneous questions. However, these questions are essential to the simulation in that novices will be asking roleplayers these questions.

To simulate a novice's route:

1. For each selected question, the system finds the locations in the all the log files where the question (or a semantically identical question) was used
2. It normalizes each of those values according to how far through the interaction the question was used. Normalizing these values is important because the interviews range from 5 minutes to 90 minutes.
3. Last, the system calculates the average of the normalized values.

This method correctly sorts questions asked at the beginning and end of the interaction. For example, greetings and eliciting the chief complaint usually happen near the beginning of a medical interview, and medical advice and exit phrases happen at the end. In Figure 5-3 we see that the list of automatically ordered questions is in roughly the correct order. Using the arrows on the left side of the interface the domain expert can re-arrange any out of order questions. "Questions to Ask" are the questions the VH will ask, in the order it will ask them. "Character Responses" are the original question-answering VH's responses to the questions.

In the scenario for the pilot study, this method automatically arranged questions in an order roughly similar to a real medical student. The Roleplay Trainer Creator provided an interface to the expert for validating and fixing the order. Reordering and validating the order of the 71 questions in the test-bed scenario required approximately 30 minutes.

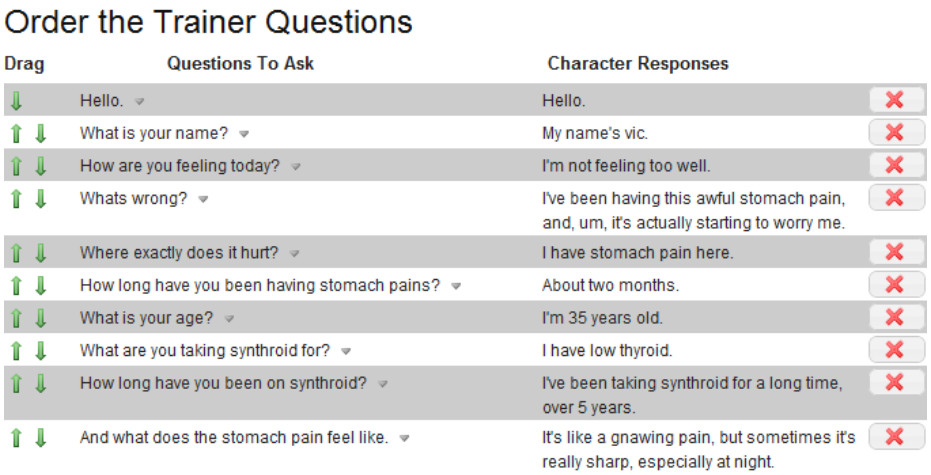

Figure 5-3. The Roleplay Trainer Creator interface for viewing automatically ordered interview questions and reordering the questions

5.2.2.3 Simulating the roleplay partner

As described in section 5.1, there are four main goals for the interaction experience. The goals are to provide roleplay trainees with:

1. Coverage of the questions that novices will ask,
2. The correct responses to those questions,
3. Experience answering the questions, and
4. Feedback on responses to those questions.

These four goals were implemented using the Virtual People Factory interaction interface. Virtual People Factory provides a web browser interface to interact with VHs using text-based chat. This interface is similar to online chat interfaces such as AOL Instant Messenger or Gmail Chat in which a user types an input, and receives a text-based response. For simulating the roleplay interaction, the Virtual People Factory interface was augmented with additional features to enable effective training of roleplay partners. A screenshot of the interaction interface is shown in Figure 5-4. The virtual roleplay partner (virtual medical student) asks questions, and the user (standardized patient) enters text based responses. If the user does not know the answer to a question, they can mouse over the “See Correct Answers” box. The system will then

display the answer that the question-answering VH would give to the current question asked by the question-asking virtual roleplay trainer. As the user goes through the experience, they can either refer back to the printed material, or use the “See Correct Answers” to learn the correct response for the case. This experience provides a way to practice formulating responses and learn the case beyond rote memorization.

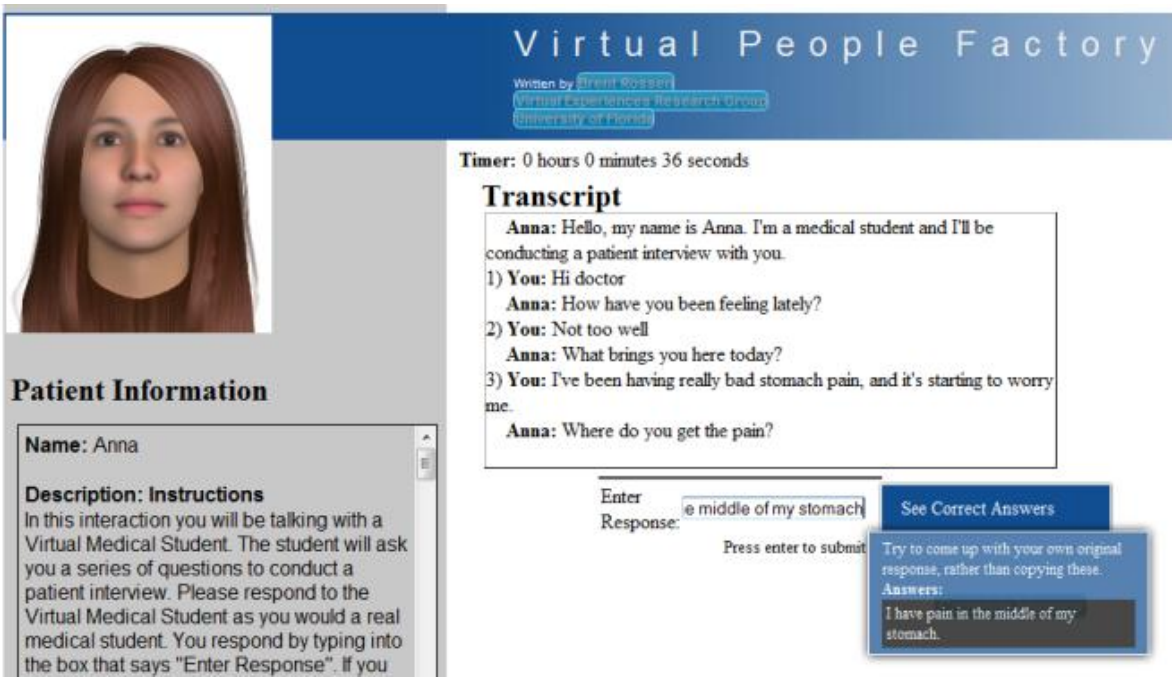

Figure 5-4. The Roleplay Trainer virtual medical student interaction interface

5.3 Pilot Study

To evaluate The Roleplay Trainer Creator we used a previously generated question-answering VH patient case. The case selected is the Dyspepsia case described in the quote below.

> You are presenting at an outpatient clinic with pain in the middle of your stomach. The onset of the pain was two months ago. You have been taking Tums twice a day for the stomach pain. You have also been having low back pain for the past three months. You have been taking two Aspirin three times a day for the back pain. The Aspirin consumption has caused a stomach ulcer, which is the source of the pain.

The case description is the same for the VH patient as it is for the standardized patients. The scenario was originally developed as a 35-year-old male, and was rewritten to be generic for use with male and female users of varying ages. This case was selected because the VH patient conversation was well developed for answering questions (Rossen et al., 2009), and it was a scenario that was not currently available to the standardized patients at the medical school where the pilot study was conducted.

The evaluation study examined the Roleplay Trainer for:

1. **Usability:** as measured by the standardized patients perceived ease of use
2. **Acceptability:** as measured by the standardized patients perceived usefulness
3. **Learning:** as measured by the standardized patients perceived feelings of preparation and confidence for playing the role of the patient pre- and post-experience

The usability and usefulness survey was based on Davis's perceived usefulness, ease of use, and user acceptance of information technology survey (Davis, 1989). There are 12 questions in this questionnaire; responses are rated from 1 (unlikely) to 7 (likely). Learning was measured by the standardized patient's perceived feelings of preparation and confidence for playing the role of the patient pre- and post- experience. Preparedness and confidence were rated on a Likert scale from 1 (not very) to 7 (very). This pilot study is intended as a feasibility study, rather than an exhaustive examination of a finished application. We further used the study to elicit feedback from standardized patients on future directions for the project.

5.3.1 Population

Five (n = 5) standardized patients at The University of Florida conducted a medical interview with the virtual medical student. The participants varied in age from 22 to 79 with a mean age of 38, and two of the participants were female. Participants varied in

previous standardized patient experience with as few as 6 previous interactions with medical students, and as many as 100 with a mean of 33.4. Participants were paid their regular hourly wage for the hour they participated in the study.

5.3.2 Procedure

1. Participants were introduced to the scenario using a three page printed description on hard-copy paper. The paper scenario provided only the most basic information about the scenario. Participants had 10 minutes to review the material.
2. They read and signed the consent form.
3. They were issued the digital pre-study questionnaire on a Dell D400 laptop using Internet Explorer 7.
4. They used the laptop to conduct the medical interview with the virtual medical student.
5. They filled out the digital post-study questionnaire

5.3.3 Metrics

Pre-study questionnaire. Collected the participant's demographics, previous experience in playing a standardized patient, and assessed their feelings of preparedness and confidence in playing the role of the patient in the given scenario. Preparedness and confidence were rated on a Likert scale from 1 - not very to 7 - very.

Post-study questionnaire. Re-assessed the participant's feelings of preparedness and confidence in playing the role of the patient. This questionnaire also assessed the participant's subjective opinion of how useful the system would be in preparing for a standardized patient case as well as the system's ease of use.

5.3.4 Results

Three out of five participant's preparedness and confidence increased from pre-survey to post-survey (see Figure 5-5). These results display a reduction of variance in participant's feelings of preparedness and confidence in displaying the patient's

symptoms. Participants who felt unprepared and low-confidence prior to the training experience felt more prepared and confident afterwards, and the participant who felt prepared and confident prior to the interaction felt less prepared and confident after.

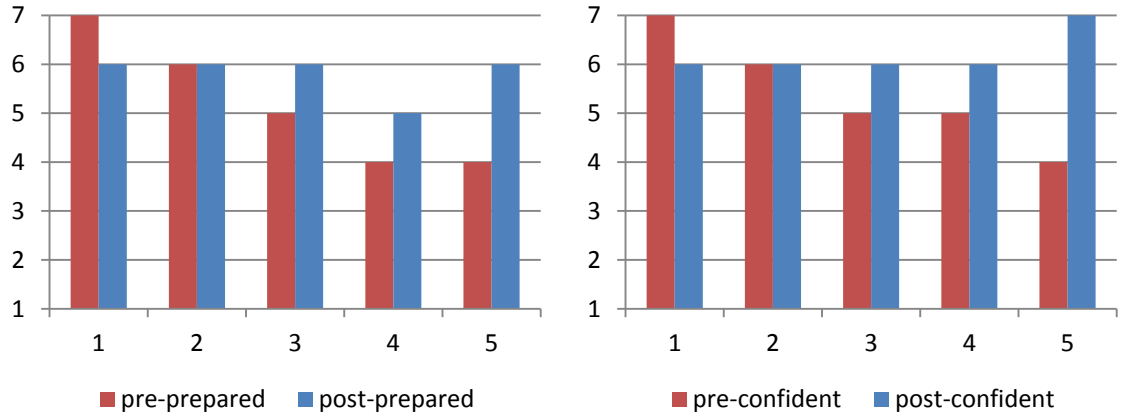

Figure 5-5. Survey results of participant's self-assessed pre- post- preparedness and pre- post- confidence

The perceived usefulness and ease of use were analyzed by computing the mean value of the perceived usefulness and ease of use questions for each participant, then taking the mean over all participants. Survey results indicate that, on average, participants felt the system was easy to use (6.3 out of 7 ± 0.9), and that it would be moderately helpful for doing their job (4.5 out of 7 ± 1.7). There was significant variance indicated by the high standard deviation; three users felt that interaction with virtual medical students would be helpful to very helpful for preparation (>= 5), one felt it would be moderately helpful (4), while one user felt the system would not be helpful (3).

5.3.5 Discussion

The results of the study indicate that training standardized patients with a virtual medical student has potential. Three out of five participants reported an increased feeling of preparedness and confidence after interacting with the virtual medical student. The one user whose ratings decreased pre- to post- survey rated himself pre-

experience at maximum (7), and so may have overrated his preparedness and confidence in the pre-survey.

According to the usability survey results, the users felt it was easy to converse with the virtual medical student. The usefulness score had a high standard deviation, indicating that some users found the system to be useful for preparing and others did not. In future studies, with a larger number of users, we may be able to determine which users gain greater benefit from the experience, such as less experienced users or users with greater comfort with technology.

5.4 Interpersonal Training Applications

We present Roleplay Trainer Creator, a system that uses logs from user interactions with a VH to generate a new unique VH. This VH is a virtual representation of the aggregate set of users. This method of generating a new VH from existing interaction logs has significant potential for interpersonal training applications. This research has shown that a question-asking VH generated using this method could become a beneficial training tool for increasing the standardization of roleplay partners. Since the time of this research, an independent study was conducted comparing the Roleplay Trainer interactions with traditional standardized patient training techniques using videos. The researchers found similar training benefit from each method, and concluded that the Roleplay Trainer has advantages in diversity, participant engagement, and availability (Palathinkal, 2011).

CHAPTER 6
CONCLUSIONS

6.1 Review of Results

In this dissertation, we claimed that:

> The proposed conversational modeling methodologies improve the efficiency of the human effort used to model a conversation. Improved efficiency results in a significantly shorter time to produce more accurate conversational models than previous methods. This reduction in time and effort enhances the applicability of virtual humans to real-world interpersonal skills education.

The results of our evaluations demonstrate that Human-centered Distributed Conversational Modeling and Conversational Knowledge Reuse are faster than Centralized Conversational Modeling and can result in a more comprehensive enumeration of the conversational space (Rossen et al., 2012; Rossen and Lok, 2012). These methods speed up the modeling process by enabling novice and expert users to create conversational models in a distributed fashion and reuse conversational knowledge from the creation of previous virtual humans in the creation of new virtual humans (VHs). These conversational modeling methods enable educators to create conversational models themselves, improve the efficiency of efforts used to model conversations, and result in conversational models with increased accuracy for both typed and spoken interactions.

The results also demonstrate that these methods have improved the applicability of VHs to real-world applications. Using Human-centered Distributed Conversational Modeling and Conversational Knowledge Reuse (Virtual Human Templates, Dynamic Knowledge Sharing, and Virtual Human Bootstrapping), both experts and novices created VH conversational models in 11 to 25 hours for new conversational domains. These models were of equivalent or better quality compared to those created in 200

hours using Centralized Conversational Modeling. As a result of this improved efficiency, novices can, for the first time, develop accurate conversational models for VH patient interactions within the time available during a single semester (Rossen et al., 2012). This increase in speed has enabled the application of VHs to education in pharmacy (Rossen et al., 2009), osteopathy (Surkunalingam et al., 2009), psychiatry (Foster et al., 2010a; Foster et al., 2010b), dentistry (Pileggi and Childs, 2011), patient-centered counseling (Jackson, 2010), dysphagia management (Rossen et al., 2012), and standardized patient training (Rossen et al., 2010). These findings demonstrate our method implementations have facilitated the creation of VHs for diverse medical contexts.

Lastly, we find that Virtual People Factory is an educationally valuable tool. Both students and healthcare educators report that the system is a viable method for educating healthcare students in communication skills (Filichia et al., 2010; Foster et al., 2010a; Foster et al., 2010b; Palathinkal, 2011; Peden et al., 2011; Pileggi and Childs, 2011; Rossen et al., 2010; Rossen et al., 2012; Rossen et al., 2009; Rossen and Lok, 2012; Shah et al., 2008; Shah et al., 2012; Shah et al., 2009a; Shah et al., 2009b; Surkunalingam et al., 2009). Educators find that having control over virtual human creation gives them the ability to provide focused learning experiences. Furthermore, educators who have used Virtual People Factory in classroom education are continuing to use Virtual People Factory on an on-going basis.

6.2 Real-World Usage

In August 2008, we opened Virtual People Factory to the public: http://vpf.cise.ufl.edu. As of September 2011, Virtual People Factory had 56 active users outside of our research group including VH researchers, healthcare practitioners,

psychologists, and even high-school students. Our healthcare collaborators have used Virtual People Factory to integrate VHs into healthcare curricula at the University of Florida College of Medicine, University of Florida College of Dentistry, Georgia Health Sciences University, Philadelphia College of Osteopathic Medicine, University of Central Florida College of Medicine, and University of South Florida College of Medicine. From their work, Virtual People Factory facilitated the creation of 37 VHs with over 600 questions each (a ten-fold increase over the previous rate of VH creation). Those VHs conducted over 2700 interactions consisting of more than 105,000 utterances.

Now that we have achieved this level of efficiency in the creation of VH conversational modeling, there is promise for expanding VH curricula. Our collaborators in healthcare education have already found the increased number of VHs to be useful, and are hoping to have many more VHs for interpersonal skills training. With the continued work of healthcare educators creating VH medical interactions, they may achieve the goal of providing a warehouse of diverse VHs.

6.3 Future Work

The methods presented in this dissertation have improved the efficiency of conversational modeling; however, there are opportunities to further improve the efficacy of the conversational modeling processes. The following is a list of possible future research to improve knowledge acquisition for VH conversational models.

- Motivational strategies: further examination of leaderboards to inspire competition, narratives to promote dramatic interest, and deadlines to encourage timely participation for novice interactions (Halan et al., 2010).
- Capture conversation: automatically capture knowledge from typed conversations between two humans, and use that knowledge to seed the conversational modeling process for domains in which virtual human templates have not yet been engineered.

- Conversational model clustering: use document-clustering algorithms to determine scenario similarity and indicate the applicability of Dynamic Knowledge Sharing for groups of VHs.
- Human-centered Distributed Conversational Modeling for Virtual Human Templates: use crowdsourcing and human-computation to expand existing templates.
- Generic response templates: develop a comprehensive template of generic stimuli and responses for fallback during deployment use. This would allow VHs to respond realistically to off-topic questions.
- Knowledge acquisition from roleplay training: use roleplay training as a source of knowledge to produce a variety of responses for the original VH. These responses could be tagged based on personality traits and used for specific learning goals (e.g. dealing with a verbally aggressive patient), or used to provide a variety of equivalent responses to increase the realism of the original VH.

These projects would further leverage the limited resources necessary for conversational modeling and improve interactions with VHs. But the ultimate goal of interpersonal simulation is not to provide ever more realistic interaction, it is to improve interactions with real humans. Using the methods provided in this dissertation, it is now more feasible to create VH based interpersonal skills training curriculums.

Now that we can more quickly create diverse sets of VHs, the next step is to validate them for additional practical applications. For example, after interviewing a variety of VH patients with stomach pain, are medical students better able to diagnose a human patient with stomach pain? Given a variety of both visual stimuli and conversational scenarios, can VHs be used to not only identify ethnic and cultural biases, but alleviate them as well? If these applications are found to provide significant learning, VH based interpersonal skills training could be validated for further integration into healthcare curriculums.

LIST OF REFERENCES

Aamodt, A. and E. Plaza, 1994. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 7, 39-59.

Barrows, H. S., 1987. Simulated (Standardized) Patients and Other Human Simulations. Chapel Hill, NC, Health Sciences Consortium.

Barrows, H. S., 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Academic Medicine 68 (6), 443.

Bearman, M., B. Cesnik and M. Liddell, 2001. Random comparison of 'virtual patient' models in the context of teaching clinical communication skills. Medical Education 35, 824-832.

Benedict, N., 2010. Virtual Patients and Problem-Based Learning in Advanced Therapeutics. American Journal of Pharmaceutical Education 74 (8).

Bergin, R. A. and U. G. H. Fors, 2003. Interactive simulated patient—an advanced tool for student-activated learning in medicine and healthcare. Computers & Education 40 (4), 361-376.

Cassell, J., 2001. Embodied conversational agents: representation and intelligence in user interfaces. AI magazine 22 (4), 67.

Colby, K., S. Weber and F. D. Hilf, 1971. Artificial paranoia. Artificial Intelligence 2, 1-25.

Davis, F. D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly 13 (3), 319-340.

Deaton, J. E., C. Barba, T. Santarelli, L. Rosenzweig, V. Souders, C. McCollum, J. Seip, B. W. Knerr and M. J. Singer, 2005. Virtual environment cultural training for operational readiness (VECTOR). Virtual Reality 8 (3), 156-167.

Deladisma, A., K. Johnsen, A. Raij, B. Rossen, A. Kotranza, M. Kalapurakal, S. Szlam, J. Bittner, D. Swinson, B. Lok and D. S. Lind, 2007a. Medical Student Satisfaction with Using a Virtual Patient System to Learn History-Taking and Communication Skills. Medicine Meets Virtual Reality.

Deladisma, A., D. Mack, T. Bernard, C. Oxendine, S. Szlam, P. Wagner, E. J. Kruse, B. Lok and D. S. Lind, 2007b. Virtual Patients Reduce Anxiety and Enhance Learning when Teaching Medical Student Sexual-History Taking Skills. Association for Surgical Education 2007 Surgical Education Week.

Dickerson, R., K. Johnsen, A. Raij, B. Lok, J. Hernandez, A. Stevens and D. S. Lind, 2005. Evaluating a Script-Based Approach for Simulating Patient-Doctor Interaction. SCS 2005 International Conference on Human-Computer Interface Advances for Modeling and Simulation, 79–84.

Dieker, L., M. Hynes, C. Stapleton and C. Hughes, 2007. Virtual Classrooms: STAR Simulator. New Learning Technology SALT.

Ellaway, R. and J. B. McGee. 2008. Virtual Patient Working Group. from http://www.medbiq.org/working_groups/virtual_patient/index.html.

Fall, L. H., N. B. Berman, S. Smith, C. B. White, J. C. Woodhead and A. L. Olson, 2005. Multi-institutional development and utilization of a computer-assisted learning program for the pediatrics clerkship: the CLIPP Project. Academic Medicine 80 (9), 847.

Filichia, L. A., E. Blackwelder, S. Halan, B. Rossen, B. Lok and J. C. Cendan, 2010. Using Virtual Characters to Standardize Patient Hand-Offs In Surgical Training Programs. 5th Annual Academic Surgical Congress. San Antonio, TX.

Foster, Londino, Noseworthy, Lind, Shah, Lok, Chuah and Rossen, 2010a. The Use of Interactive Virtual Patients in an Integrated Psychiatry-Neuroanatomy Course and a Psychiatry Clerkship. ADMSEP.

Foster, Noseworthy, Shah, Lind, Lok, Chuah and Rossen, 2010b. Evaluation of Medical Student Interaction with a Bipolar Virtual Patient Scenario Written by a Peer Support Specialist – a Pilot Study. ADMSEP. Jackson Hole, WY.

Freedman, J., A. Levin, C. Westbrook, P. Edwards-Daughtery, J. Hurliman, F. Taney and J. Neal. 2010. Open Metaverse Foundation - libopenmetaverse. from http://www.openmetaverse.org/.

Fukayama, A., T. Ohno, N. Mukawa, M. Sawaki and N. Hagita. 2002. Messages Embedded in Gaze of Interface Agents - Impression Management with Agent's Gaze. ACM SIGCHI.

Glass, J., E. Weinstein, S. Cyphers, J. Polifroni, G. Chung and M. Nakano, 2005. A framework for developing conversational user interfaces. Computer-Aided Design of User Interfaces IV, 349-360.

Goh, O. S., C. C. Fung, K. W. Wong and A. Depickere, 2007. Embodied conversational agents for H5N1 pandemic crisis. Journal of Advanced Computational Intelligence Vol 11 (3).

Halan, S., B. Rossen, J. Cendan and B. Lok. 2010. High Score!-Motivation Strategies for User Participation in Virtual Human Development. 10th International Conference on Intelligent Virtual Agents (IVA), Philadelphia, Pennsylvania, Springer.

Halan, S., B. Rossen, M. Crary and B. Lok, 2012. Constructionism of virtual humans to improve perceptions of conversational partners. ACM SIGCHI. In Review.

Hayes, K. A. and C. U. Lehmann, 1996. The interactive patient: a multimedia interactive educational tool on the World Wide Web. MD Comput 13 (4), 330-4.

Hill, R., J. Gratch, S. Marsella, J. Rickel, W. Swartout and D. Traum, 2003. Virtual Humans in the Mission Rehearsal Exercise System. Künstliche Intelligenz 4 (03), 5-10.

Howe, J., 2006. The rise of crowdsourcing. Wired magazine 14 (6), 1-4.

Huang, G., R. Reynolds and C. Candler, 2007. Virtual patient simulation at US and Canadian medical schools. Academic Medicine 82 (5), 446.

Hubal, R. C. and R. S. Day, 2006. Informed consent procedures: An experimental test using a virtual character in a dialog systems training application. Journal of Biomedical Informatics.

Hubal, R. C., P. N. Kizakevich, C. I. Guinn, K. D. Merino and S. L. West, 2000. The virtual standardized patient. Simulated patient-practitioner dialog for patient interview training. Stud Health Technol Inform 70, 133-8.

Iacobelli, F. and J. Cassell, 2007. Ethnic Identity and Engagement in Embodied Conversational Agents. Intelligent Virtual Agents (IVA). Paris, France, Springer.

Itin, C., 1999. Reasserting the philosophy of experiential education as a vehicle for change in the 21st century. Journal of Experiential Education 22 (2), 91-98.

Jackson, J., 2010. USF Health - Virtual Standardized Patient Simulation (Video). from http://vimeo.com/16291272.

Jenkins, V. and L. Fallowfield, 2002. Can Communication Skills Training Alter Physicians' Beliefs and Behavior in Clinics? Journal of Clinical Oncology 20 (3), 765-769.

Johnsen, K., 2008. Design and Validation of a Virtual Human System for Interpersonal Skills Education. Computer Information Science and Engineering. Gainesville, FL, University of Florida. Doctor of Philosophy, 146.

Johnsen, K., R. Dickerson, A. Raij, C. Harrison, B. Lok, A. Stevens and D. S. Lind, 2006. Evolving an Immersive Medical Communication Skills Trainer. Presence: Teleoperators & Virtual Environments 15 (1), 33-46.

Johnsen, K., R. Dickerson, A. Raij, B. Lok, J. Jackson, M. Shin, J. Hernandez, A. Stevens and D. S. Lind. 2005. Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills. IEEE Virtual Reality.

Johnsen, K., A. Raij, A. Stevens, D. S. Lind and B. Lok, 2007. The Validity of a Virtual Human System for Interpersonal Skills Education. ACM SIGCHI.

Kamin, C., R. Deterding and M. A. Y. Lowry, 2002. Student's perceptions of a virtual PBL experience. Academic Medicine 77 (11), 1161.

Kenny, P., A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella and D. Piepol. 2007. Building interactive virtual humans for training environments. ITSEC, NTSA.

Kenny, P., T. D. Parsons, J. Gratch and A. A. Rizzo, 2008. Evaluation of Justina: A Virtual Patient with PTSD. Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008, Proceedings.

Kotranza, A., 2009. Mixed Reality Interpersonal Simulation Affords Cognitive, Psychomotor, and Affective Learning. Computer Information Sciences and Engineering. Gainesville, FL, University of Florida. Doctor of Philosophy, 286.

Kotranza, A., K. Johnsen, J. Cendan, B. Miller, D. S. Lind and B. Lok. 2009a. Virtual multi-tools for hand and tool-based interaction with life-size virtual human agents, IEEE.

Kotranza, A. and B. Lok. 2008. Virtual Human + Tangible Interface = Mixed Reality Human. An Initial Exploration with a Virtual Breast Exam Patient. IEEE Virtual Reality.

Kotranza, A., B. Lok, A. Deladisma, C. Pugh and D. S. Lind, 2009b. Mixed reality humans: Evaluating behavior, usability, and acceptability. IEEE Transactions on Visualization and Computer Graphics 15 (3), 369-382.

Lei, G. A. O., L. I. Xiao-song, W. A. N. Xue-hong and Z. Cheng, 2004. An application of multivariate generalizability theory in the study on the reliability of objective structured clinical examination [J]. China Higher Medical Education 3.

Leong, S. L., C. D. Baldwin and A. M. Adelman, 2003. Integrating Web-based computer cases into a required clerkship: development and evaluation. Academic Medicine 78 (3), 295.

Leuski, A., R. Patel, D. Traum and B. Kennedy, 2006. Building effective question answering characters. Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue.

Leuski, A. and D. Traum. 2010. Practical language processing for virtual humans. Twenty-Second Annual Conference on Innovative Applications of Artificial Intelligence.

Lidwell, W., K. Holden and J. Butler, 2003. Universal principles of design, Rockport.

Lind, D. S. and B. Lok, 2006. The Role of Virtual Patients in Medical Education: Teaching Tool Versus Technological Trend. Focus on Surgical Education 2006(1), 23.

Manganas, A., M. Tsiknakis, E. Leisch, M. Ponder, T. Molet, B. Herbelin, N. Magnenat-Thalmann, D. Thalmann, M. Fato and A. Schenone, 2004. The Just Vr Tool: An Innovative Approach to Training Personnel for Emergency Situations Using Virtual Reality Techniques. The Journal on Information Technology in Healthcare 2 (399-412).

Mauldin, M. L. 1994. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition, JOHN WILEY & SONS LTD.

Nutter, D. and M. Whitcomb, 2001. The AAMC Project on the Clinical Education of Medical Students. Association of American Medical Colleges. Washington, DC.

Palathinkal, J., 2011. The Effectiveness of Virtual Humans vs. Pre-Recorded Humans in a Standardized Patient Performance Assessment. College of Sciences. Orlando, University of Central Florida. Doctor of Philosophy in Modeling and Simulation.

Parsons, T. D., P. Kenny, C. A. Ntuen, C. S. Pataki, M. T. Pato, A. A. Rizzo, C. St-George and J. Sugar, 2008. Objective structured clinical interview training using a virtual human patient. Studies in Health Technology and Informatics 132, 357.

Peden, M., J. Chuah, A. Kotranza, K. Johnsen, B. Lok and J. Cendan, 2011. NERVE- A Three Dimensional Patient Simulation for Evaluating Cranial Nerve Function. MedEdPORTAL,

Pileggi, R. and G. S. Childs, 2011. Virtual Humans in Dental Education: The case of an emergency patient American Dental Education Association.

Plant, E. A., B. M. Peruche and D. A. Butz, 2005. Eliminating Automatic Racial Bias: Making Race Non-Diagnostic for Responses to Criminal Suspects. Journal of Experimental Social Psychology 41, 141-156.

Raij, A., 2009. Using Immersion and Information Visualization to Analyze Human-Virtual Human Interactions. Computer Information Sciences and Engineering. Gainesville, FL, University of Florida. Doctor of Philosophy, 205.

Raij, A., K. Johnsen, R. Dickerson, B. Lok, M. Cohen, A. Stevens, T. Bernard, C. Oxendine, P. Wagner and D. S. Lind, 2007. Comparing Interpersonal Interactions with a Virtual Human to those with a Real Human. IEEE Transactions on Visualization and Computer Graphics 13 (3), 443-457.

Raij, A., A. Kotranza, D. S. Lind and B. Lok, 2009. Virtual experiences for social perspective-taking. IEEE Virtual Reality. Lafayette, LA.

Raij, A. and B. Lok. 2008. IPSViz: An After-Action Review Tool for Human-Virtual Human Experiences. IEEE Virtual Reality 2008.

Reiter, E., S. Sripada and R. Robertson, 2003. Acquiring correct knowledge for natural language generation. Journal of Artificial Intelligence Research 18 (491-516).

Rizzo, A. S., T. Parsons, J. G. Buckwalter and P. Kenny, 2010. A new generation of intelligent virtual patients for clinical training. IEEE Virtual Reality.

Roque, A. and D. Traum. 2007. A model of compliance and emotion for potentially adversarial dialogue agents. SIGDial Workshop on Discourse and Dialogue, Citeseer.

Rossen, B., 2010. MySQL Ajax Database Access Layer (MADAL). from http://code.google.com/p/madal/.

Rossen, B., J. Cendan and B. Lok. 2010. Using virtual humans to bootstrap the creation of other virtual humans. The 10th International Conference on Intelligent Virtual Agents, Springer.

Rossen, B., A. Deladisma, D. Lind and B. Lok, 2008. Virtual Humans Elicit Skin-Tone Bias Consistent with Real-World Skin-Tone Biases. International Conference on Intelligent Virtual Agents. Tokyo, Japan.

Rossen, B., S. Halan and B. Lok, 2012. Reusing Conversational Data to Rapidly Generate Virtual Human Conversational Models. Journal of Autonomous Agents and Multi-Agent Systems In Review.

Rossen, B., D. S. Lind and B. Lok, 2009. Human-centered Distributed Conversational Modeling: Efficient Modeling of Robust Virtual Human Conversations. International Conference on Intelligent Virtual Agents. Amsterdam, Netherlands.

Rossen, B. and B. Lok, 2012. A Crowdsourcing Method to Develop Virtual Human Conversational Agents. International Journal of Human-Computer Studies Accepted (awaiting publication).

Ruttkay, Z., E. Andre, W. L. Johnson and C. Pelachaud, 2004. Evaluating Embodied Conversational Agents. Evaluating Embodied Conversational Agents 4121,

Saleh, N., 2010. The Value of Virtual Patients in Medical Education. Annals of Behavioral Science and Medical Education 16 (2), 29-31.

Sebe, N., 2010. Human-centered Computing. Handbook of Ambient Intelligence and Smart Environments, 349-370.

Shah, H., P. Fox, B. Rossen, A. Deladisma, A. Gehlot, T. Lord, A. Gucwa, A. Mouna, J. Beatty, B. Rosson, B. Lok and D. Lind, 2008. A Pilot Study to Investigate the Efficacy of a Novel Interactive Web-Based Virtual Clinical Scenario System (Virtual People Factory) in Medical Education. Society for Simulation in Healthcare.

Shah, H., B. Rossen and A. Foster, 2012. Interactive Virtual Patient Scenarios: an Evolving Tool in Psychiatric Education. Academic Psychiatry Accepted, Awaiting Publication.

Shah, H., B. Rossen, D. Lind and B. Lok, 2009a. Pilot Study to Evaluate the Use of an Online Virtual Patient System to Teach Interviewing Skills to First-Year Medical Students. National Conference of Family Medicine Residents and Medical Students. Kansas City.

Shah, H., B. Rossen, B. Lok, A. House, D. S. Lind and A. Foster, 2009b. A Pilot Study to Evaluate the Use of an Interactive Virtual Patient with Depression to Teach History-Taking Skills in a Psychiatry Clerkship. ADMSEP. Portsmouth, NH.

Shortliffe, E. H., 1976. Computer-Based Medical Consultations: MYCIN. New York.

Singh, P., T. Lin, E. T. Mueller, G. Lim, T. Perkins and W. Li Zhu, 2002. Open Mind Common Sense: Knowledge Acquisition from the General Public. On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE, 1223-1237.

Stevens, A., M. Cohen, K. Johnsen, R. Dickerson, A. Raij, R. Wells, C. Oxendine, P. Wagner, T. Bernard, J. Cendan, M. Duerson, R. pauly, B. Lok and D. S. Lind. 2006. Implementing a Virtual patient into the Medical School Curriculum at the University of Florida. Southern Group on Educational Affairs.

Stevens, A., J. Hernandez, K. Johnsen, R. Dickerson, A. Raij, J. Jackson, M. Shin, J. Cendan, R. Watson, M. Duerson, B. Lok, M. Cohen, P. Wagner and D. S. Lind, 2005. The Use of Virtual Patients to Teach Medical Students Communication Skills. American Journal of Surgery 191 (6), 806-811.

Surkunalingam, N., J. Walker, C. Taranto, S. McCaskill, C. Blanchard, D. S. Lind, B. Lok and B. Rossen, 2009. A Study to Investigate the Efficacy of a Novel Interactive Web-Based Virtual Clinical Scenario System (Virtual People Factory) in Medical Education. American Osteopathic Association National Conference.

Sutton, S., D. G. Novick, R. Cole, P. Vermeulen, J. de Villiers, J. Schalkwyk and M. Fanty. 1996. Building 10,000 spoken dialogue systems. International Conference on Spoken Language Processing, IEEE.

Tamblyn, R. M., D. J. Klass, G. K. Schnabl and M. L. Kopelow, 2009. The accuracy of standardized patient presentation. Medical Education 25 (2), 100-109.

Traum, D., 2008. Talking to Virtual Humans: Dialogue Models and Methodologies for Embodied Conversational Agents. LECTURE NOTES IN COMPUTER SCIENCE 4930, 296.

Triola, M. M., N. Campion, J. B. McGee, S. Albright, P. Greene, V. Smothers and R. Ellaway. 2007. An XML standard for virtual patients: exchanging case-based simulations in medical education. AMIA Annual Symposium, American Medical Informatics Association.

Ullrich, S., K. Bruegmann, H. Prendinger and M. Ishizuka, 2008. Extending MPML3D to Second Life. Intelligent Virtual Agents. Tokyo, Japan. 5208, 281-288.

Villaume, W. A., B. A. Berger and B. N. Barker, 2006. Learning Motivational Interviewing: Scripting a Virtual Patient. American Journal of Pharmaceutical Education 70 (2).

Vincent, C., M. Young and A. Phillips, 1995. Why Do People Sue Doctors? A Study of Patients and Relatives Taking Legal Action. Obstetrical and Gynecological Survey 50, 103-105.

von Ahn, L. and L. Dabbish, 2004. Labeling images with a computer game. Proceedings of the SIGCHI conference on Human factors in computing systems, 319-326.

Wagner, W. P. 1990. Issues in knowledge acquisition. Proceedings of the ACM SIGBDP conference on Trends and directions in expert systems ACM.

Wallace, R. S., 2005. Be your own botmaster: The step by step guide to creating, hosting and selling your own AI chatbot on Pandorabots. ALICE AI Foundation.

Walters, K., D. Osborn and P. Raven, 2005. The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. Medical Education 39 (3), 292-298.

Weizenbaum, J., 1966. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM 9 (1), 36-45.

Wessel, J., R. Williams, E. Finch and M. Gemus, 2003. Reliability and validity of an objective structured clinical examination for physical therapy students. Journal of Allied Health 32 (4), 266-269.

Westberg, J. and H. Jason, 2001. Fostering reflection and providing feedback, Springer.

Yedidia, M. J. and M. Lipkin, 2003. Effect of Communications Training on Medical Student Performance. JAMA 290, 1157-1165.

Zanbaka, C. A., A. C. Ulinski, P. Goolkasian and L. F. Hodges, 2007. Social responses to virtual humans: implications for future interface design. Proceedings of the SIGCHI conference on Human factors in computing systems, 1561-1570.

BIOGRAPHICAL SKETCH

Brent H. Rossen was born in 1981 in Petaluma, California, to Jan and Joel Rossen. Brent was raised in California until the age of 13, and then moved to Coral Springs, Florida. In Coral Springs, he graduated with honors from J.P. Taravella High in 2000. That year, he received a full scholarship to attend the University of Florida. In 2004, Brent was a University Scholar advised by Dr. Paul Fishwick. In 2005, he graduated Summa Cum Laude with his B.S. in Digital Arts and Sciences.

He continued at the University of Florida to pursue his Ph.D. in computer engineering under the supervision of Dr. Benjamin Lok, and received his Ph.D. in the fall of 2011. Brent was awarded a UF Alumni Fellowship to support his research, which focuses on conversational modeling, virtual human interfaces, and their emerging use for interpersonal skills training. His Ph.D. work explored novel approaches to creating conversational models for simulating natural language conversations with virtual humans. His work received significant international recognition in both the fields of computer science and medicine with 17 publications in journals and conferences, including a journal article in the International Journal of Human Computer Studies. Brent and his collaborators have applied for patents on the technology described in his dissertation.

On June 20, 2009, Brent married his college sweetheart, Elizabeth. Brent and Elizabeth enjoy traveling the world together – they visited five countries during graduate school, and plan to see many more. After a summer internship, Brent received a position at Microsoft working on massively scalable storage systems. Brent and Elizabeth currently reside in Redmond, Washington with their dogs Brandy and Beijing.