Available online at www.sciencedirect.com

SciVerse ScienceDirect

Int. J. Human-Computer Studies 70 (2012) 301–319

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

A crowdsourcing method to develop virtual human conversational agents

Brent Rossen*, Benjamin Lok

Computer & Information Sciences & Engineering, University of Florida, Gainesville, FL 32611, USA

Received 17 March 2011; received in revised form 24 September 2011; accepted 30 November 2011
Communicated by: C. Bartneck
Available online 5 January 2012

Abstract

Educators in medicine, psychology, and the military want to provide their students with interpersonal skills practice. Virtual humans offer structured learning of interview skills, can facilitate learning about unusual conditions, and are always available. 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 (educators). We address these problems using a novel crowdsourcing method entitled Human-centered Distributed Conversational Modeling. This method facilitates collaborative development of virtual humans by two groups of end-users: domain experts (educators) and domain novices (students). We implemented this method in a web-based authoring tool called Virtual People Factory. Using Virtual People Factory, medical and pharmacy educators are now creating natural language virtual patient interactions on their own. This article presents the theoretical background for Human-centered Distributed Conversational Modeling, the implementation of the Virtual People Factory authoring tool, and five case studies showing that Human-centered Distributed Conversational Modeling has addressed the logistical cost for acquiring knowledge.
Published by Elsevier Ltd.

Keywords: Virtual humans; Agents and intelligent systems; Human-centered computing; Distributed knowledge acquisition; End-user programming

1. Introduction

Virtual humans (VHs) for natural language conversation are increasingly popular for communication skills training. Projects in medicine (Dickerson et al., 2005; Villaume et al., 2006), psychology (Kenny et al., 2008), and the military (Kenny et al., 2007) have been created with collaborative effort between end-users (educators and students), artists, programmers, and conversation knowledge engineers (generally computer scientists). However, these collaborators find creating the necessary conversational corpora to be logistically difficult and time consuming (Dickerson et al., 2005; Glass et al., 2005; Kenny et al., 2007, 2008; Villaume et al., 2006). Preparing a virtual human (VH) to conduct a free-form conversation can take hundreds of hours over several months. For example, Vic is a VH who plays the role of a patient having stomach pain. Vic was created to be capable of a 10 min free-form conversation about his symptoms with a pharmacy student. To obtain 75% accuracy in Vic's conversational model required 200 h of knowledge engineer and educator time dispersed over six months. The labor intense nature of development restricts the number of VHs that can be created. As a result of such extensive domain expert and knowledge engineer time requirements, we are unable to create the number of scenarios needed to implement an interpersonal skills training curriculum (Triola et al., 2007). To address these problems, we propose a novel crowdsourcing method for developing conversational models, Human-centered Distributed Conversational Modeling (HDCM). Using HDCM results in a more complete and accurate conversational model in a significantly shorter time.

The proposed method addresses the modeling of natural language conversations for interpersonal skills education.

*Corresponding author. Fax: +1 352 392 1220.
E-mail addresses: brossen@cise.ufl.edu (B. Rossen), lok@cise.ufl.edu (B. Lok).

1071-5819/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.ijhcs.2011.11.004

Vic is one of several VHs created in the last few years for domain-specific conversations. These 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, 2008; Leuski et al., 2006). Natural language conversations are important for testing and facilitating skills transfer to real interactions (Johnsen et al., 2007). To develop robust natural language conversational models, knowledge engineers must acquire a large 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, 2008; Leuski et al., 2006; Reiter et al., 2003).

Typical approaches to creating conversational corpora include – recordings of people in "natural" or staged interactions, asking experts, and Wizard of Oz (human-controlled) interactions (Ruttkay et al., 2004). The usual approach of conversational knowledge engineers is to gather starting stimuli from these resources, refine with user interactions, validate with experts, and repeat (Dickerson et al., 2005; Kenny et al., 2007, 2008; Leuski et al., 2006). This process follows the patterns established by expert systems in which a knowledge engineer is required for translating domain knowledge into machine-readable information (Shortliffe, 1976).

We will hereafter refer to this method as Centralized Conversational Modeling (CCM) because of the knowledge engineer's role as the hub for transferring information from experts and novices (students) to the conversation corpus (conversational knowledge database) as shown in Fig. 1. As indicated above, the CCM process is slow, and educators report that communication bottlenecks are often frustrating.

In contrast, the HDCM approach proposes that VH users (as opposed to knowledge engineers) generate the model themselves through a distributed (web-based) system. Using HDCM, domain experts and novices collaborate to teach the VH how to converse. HDCM results in a flow of data, Fig. 2, which does not have the knowledge engineer bottleneck as in CCM, Fig. 1. Domain novices speak with the VH, which gathers new stimuli, and the domain experts add new responses to these stimuli. Effectively, the novices and the expert are collaborating to teach the VH how to conduct this domain-specific conversation.

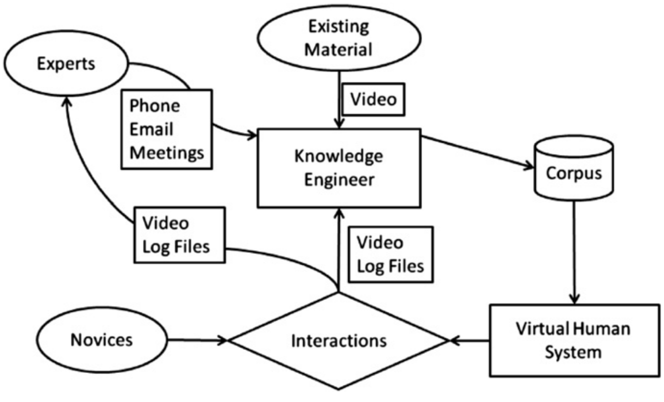


Fig. 1. The flow of data in CCM.

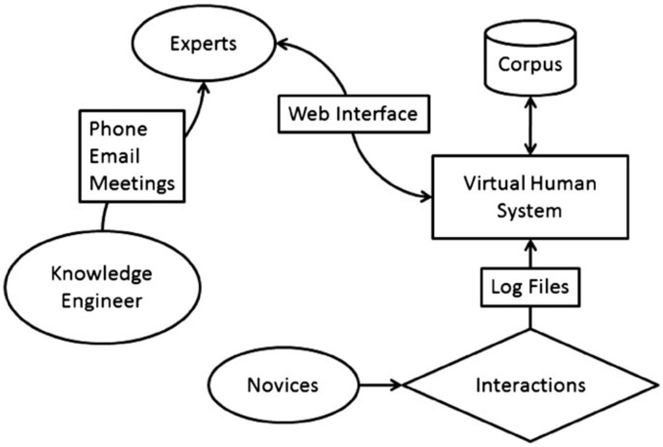


Fig. 2. The flow of data in HDCM.

HDCM applies the ideas of Crowdsourcing (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 HDCM process is driven by the domain experts, thus allowing the experts to drive the process of creating VH based training curricula and to focus the material on learning goals.

2. Background

2.1. Creating conversational corpora

The required size of a corpus to facilitate robust conversations hinders the use of VH applications in communication skills education. These conversations require a large amount of knowledge about potential speech inputs and outputs – thousands of stimuli associated with hundreds of responses. The logistical difficulties in knowledge acquisition for natural language generation have been explored by Reiter (Reiter et al., 2003). Reiter identified the problems with directly asking experts for knowledge. Expert knowledge is a good "starting point", but is not detailed enough for generalization. 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. HDCM incorporates these ideas by using a collaborative authoring tool that collects knowledge from novice users, which is validated by the expert users.

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 to 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 we take 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 our case, that side effect is teaching a VH to conduct an interaction.

2.2. *Virtual human training applications*

Currently, three primary application fields of VHs are: military, psychology, and medicine.

In military simulation, VHs are generally combatants, civilians, and fellow team members. The University of Southern California's Institute for Creative Technologies has applied VH technology to military leadership training in its Mission Rehearsal Exercise and SASO trainers (Hill et al., 2003; Kenny et al., 2007; Traum, 2008). This system allows users to interact naturally through speech with life-size VHs.

In psychology, Justina, a VH with post-traumatic stress disorder, gives clinical psychologists an opportunity to practice clinical interview skills through spoken interactions (Kenny et al., 2008). VHs in psychology are also used as spectators for practicing public speaking (Slater et al., 1999).

In medicine, VHs are often used to represent virtual patients. VH patients simulate lifelike clinical scenarios in which the user becomes the health care professional making clinical decisions. VH 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 a rapport with the patient through empathy, professionalism, and proper procedure. In surveying the existing systems (Benedict, 2010; Bergin and Fors, 2003; Ellaway and McGee, 2008; Hubal et al., 2000; Johnsen et al., 2007; Kenny et al., 2008), healthcare students interact with VH patients by: typing, choosing from a predefined list, or speaking utterances to the patient (e.g., "how long have you had the pain?"). The VH patient is represented as either recorded video of actors, animated video, still images, or a 2D or 3D animated virtual character. The VH patient is displayed on a monitor, in a web-browser, or at life-size with a projector, large screen TV, or head-mounted display. In this paper, we focus on improving typed and spoken conversations with VH patients, regardless of representation or presentation.

The University of Florida's Virtual Experiences Research group, the Interpersonal Simulator allows medical students to practice interview skills through natural interaction with life-size VHs (Dickerson et al., 2005). Interaction with these VHs has been shown to reduce anxiety and enhance learning for taking a patient's sexual history (Deladisma et al., 2007). The Research Triangle Institute has also created a virtual standardized patient to interact with medical practitioners for practicing patient assessment, diagnosis, and interviewing skills (Hubal et al., 2000).

The collective goals of these research efforts are 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., 2007; Iacobelli and Cassell, 2007; Zanbaka et al., 2007), or to construct cognitively accurate dialog modeling architectures (Ellaway and McGee, 2008; Traum, 2008). Effective interaction has progressed, but speed and independent function of domain experts remains problematic. Our solution focuses on making these VH technologies more practical. With rapid and accurate knowledge acquisition, we may be able to increase the potential for widespread adoption of VHs as training tools.

2.3. *Virtual humans for healthcare training*

The current work focuses on the field of healthcare interview training. Patient interviewing skills are crucial in all areas of healthcare. Healthcare students traditionally train for these interviews by interacting with real patients in various healthcare environments. Given the importance of accurate interview results, there is demand for earlier practice opportunities. Healthcare students should have ample practice before talking to their first real patient. For this reason, healthcare students also interact with standardized patients. 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). This is why educators in the healthcare fields want to provide VH patients for practicing procedural, diagnosis, and communication skills before students interact with standardized patients.

The creation of a VH patient requires significant collaborative effort between healthcare educators and computer scientists. As computer scientists, we rarely have the knowledge to create healthcare scenarios. Healthcare educators have that knowledge; however, educators do not have the computer science expertise necessary to create VH patients capable of natural language conversations. A system that enables educators to construct VH patients for healthcare interview training without computer science expertise would enhance development efficiency and expand utilization. In the implementation section (Section 4) we discuss our process for enabling healthcare educators to create natural language VH patient scenarios themselves.

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 that word in it. These systems also include interfaces for ordering x-rays, labs, and other tests. These predefined list interactions focus on the fact finding mission in order to reach a diagnosis. In contrast, our system focuses on simulating a natural conversation in order to train interviewing skills. Improving interview skills requires practicing those skills, and natural interaction performs better than linear or forced branching for training communication skills (Bearman et al., 2001; Saleh, 2010; Yedidia and Lipkin, 2003).

In our system, the student interviews the VH patient and asks questions about their present health and past medical history in their own words, in any order. The healthcare student's goal is to develop skills by practicing which questions to ask, how to ask the questions, and how to empathize with the patient.

Prior to the current work, the Virtual Experiences Research Group at the University of Florida has created three VH patients using CCM. These VH patients have been shown to enhance teaching and evaluation of interviewing communication skills (Deladisma et al., 2007; Dickerson et al., 2005). In addition, our research has demonstrated that students who perform well with VH patients also perform well with standardized patients and vice versa (Johnsen et al., 2007). The success of these natural language VH patients has facilitated integration of VH patients into medical curriculums. These VH patients are used in addition to traditional training techniques using standardized patients as well as expert mentoring programs. They are implemented as preparation for standardized patient and real patient interviews. While these VH patient systems are currently useful in a limited scope, to provide wide-spread functional benefit to medical schools tens to hundreds of VH patient scenarios are necessary, resulting in unrealistic time and cost requirements (Triola et al., 2007). The creation of each VH patient with CCM required approximately 200 h over six months. The first three VH patients were created over the course of four years. The HDCM approach reduces this time significantly, thus increasing the possibility for VH patient based curricula.

2.4. Challenges in centralized conversational modeling

Three challenges hinder the creation of conversation corpuses that enumerate the stimulus-response space of a conversation.

2.4.1. Challenge 1: enough detail for generalization

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 section 5.1, 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 less than 10 of these semantically unique questions. Further, experts do not phrase questions the same way novices do. With more experience, clinicians use shorter, more focused questions and infer information from their past experiences (Westberg and Jason, 2001). 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).

2.4.2. Challenge 2: use and creation of the corpora resources

Typical approaches to creating corpora using CCM 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 – non-healthcare professionals cannot view real patient interviews. Even with staged interactions, such as having students interview 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 have the same drawbacks as using actors in terms of availability, compensation, and standardization. An additional problem is that of extracting utterances from these 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. This process consists of determining if non-responses or incorrect responses are due to missing stimuli, stimuli matching errors or speech recognition errors. These logistical issues have a time cost for both knowledge engineers and collaborators.

2.4.3. Challenge 3: knowledge engineers may not know the domain, so they must collaborate with domain experts to validate the stimuli and create new responses.

The third challenge is one of collaboration. 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 difficulties due to differing backgrounds.

In practice, these three challenges result in few iterations of user testing, and each iteration 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.

HDCM addresses these challenges. HDCM removes the manual identification of stimuli from video by knowledge engineers, and creates faster collaboration through a distributed system. Additionally, the system engages novices directly in the process of knowledge acquisition for conversational modeling.

3. 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 CCM. We see in Section 2.4, 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 of creating a 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 conversation 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 gets fewer and fewer new stimuli (see Section 5.1.2 for details).

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 the 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 5.1, for a 20 min 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 (extensible markup language) or case-based reasoning. Fig. 3 shows the iterative process end-users follow for creating a VH conversational model and is described in more detail below.

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

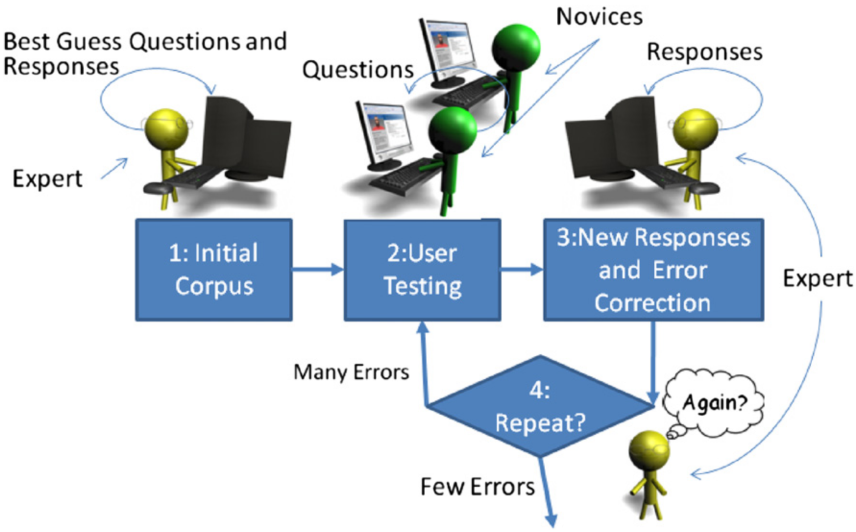


Fig. 3. HDCM design process.

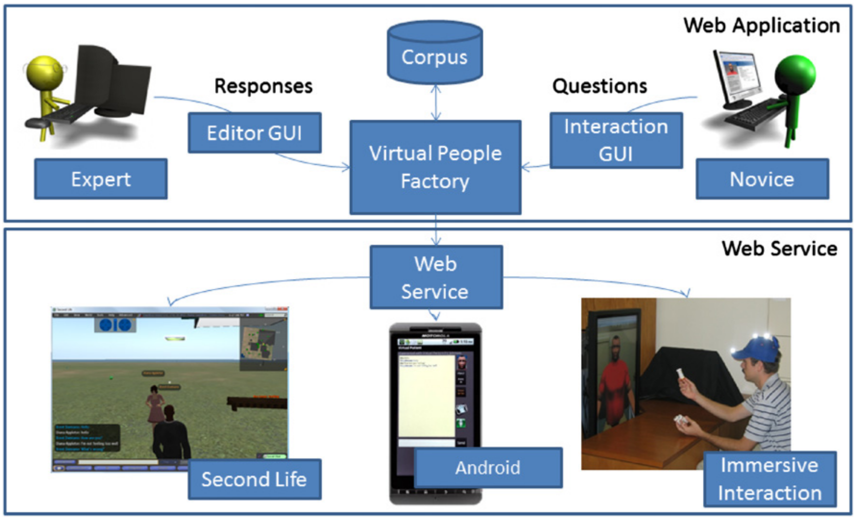


Fig. 4. Virtual People Factory system overview.

response, but the system fails to match a valid stimulus to that appropriate response. A false positive occurs when an inappropriate response is given, based on a mismatch of stimulus to an item in the corpus. These errors are entered into 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.

This HDCM method is implemented in the web-based application Virtual People Factory (VPF).

4. Virtual People Factory architecture

Virtual People Factory (VPF) is a platform for the development and deployment of VHs using the HDCM method. The VPF platform enables domain experts and novices to collaboratively create VHs. As shown in Fig. 4, 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:

(i) a browser-based interaction system,
(ii) a VH editor system, and
(iii) a web-service developer application programming interface (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 of VH scripts.

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 face to make as part of the response, as well as the timing of that face
- *Discovery:* a tag indicating an important piece of information contained in the response
- *Topic:* a tag indicating the subject or theme of the response

4.1. Virtual People Factory server

As shown in Fig. 5, VPF uses a client-server architecture. The server portion of VPF runs on a single PC containing a Core2 Duo Quad-Core processor and 4 GB 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 (MADAL). All communication is performed over http using Ajax calls. Data is marshaled and passed from one application to another in JavaScript Object Notation or XML format across http by the *VPF Web-Service API.*

4.2. VPF web-based clients

The client side runs 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). The VPF interaction and editor interfaces were implemented using html, Cascading Style Sheets, and JavaScript with jQuery.

4.2.1. 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, 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 Fig. 6 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.

4.2.1.1. 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, during interactions VPF gathers three types of

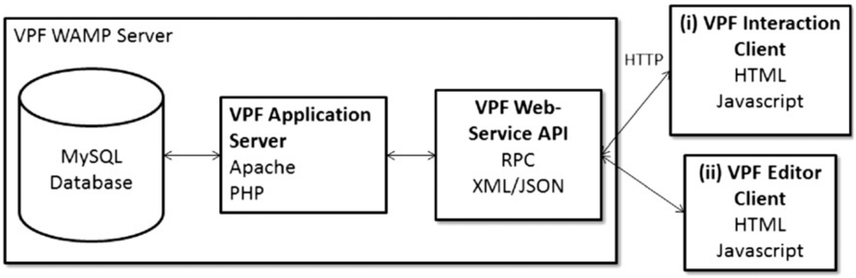
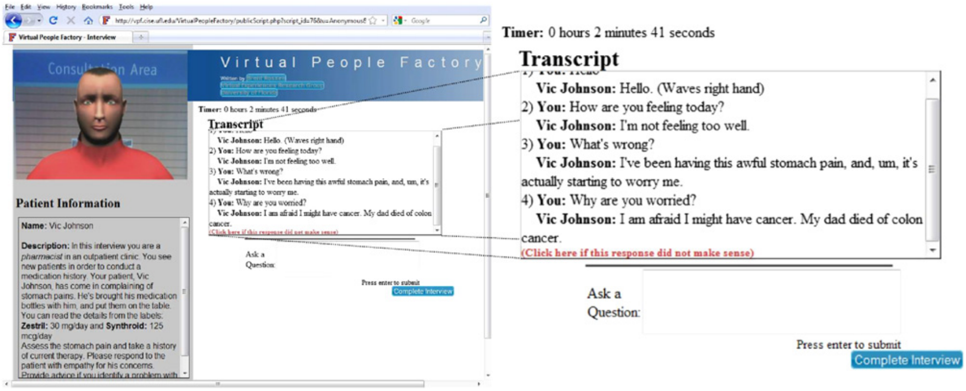


Fig. 5. The VPF architecture on Windows, Apache, MySQL, and PHP (WAMP).


File Edit View History Bookmarks Tools Help
http://vpf.cise.ufl.edu/VirtualPeopleFactory/publicScript.php?script_id=75&u=Anonymous
Virtual People Factory - Interview
Consultation Area
Virtual People Factory
Written by Brent Rossen
Virtual Experiences Research Group
University of Florida
Complete Interview

Patient Information

Name: Vic Johnson

Description: In this interview you are a *pharmacist* in an outpatient clinic. You see new patients in order to conduct a medication history. Your patient, Vic Johnson, has come in complaining of stomach pains. He's brought his medication bottles with him, and put them on the table. You can read the details from the labels: **Zestril:** 30 mg/day and **Synthroid:** 125 mcg/day
Assess the stomach pain and take a history of current therapy. Please respond to the patient with empathy for his concerns. Provide advice if you identify a problem with

Timer: 0 hours 2 minutes 41 seconds

Transcript

Vic Johnson: Hello. (Waves right hand)
2) **You:** How are you feeling today?
Vic Johnson: I'm not feeling too well.
3) **You:** What's wrong?
Vic Johnson: I've been having this awful stomach pain, and, um, it's actually starting to worry me.
4) **You:** Why are you worried?
Vic Johnson: I am afraid I might have cancer. My dad died of colon cancer.
(Click here if this response did not make sense)

Ask a Question:

Press enter to submit

Fig. 6. Browser-Based interaction GUI.

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 Fig. 6). 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.

4.2.2. 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, sending them 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 4.3 below for more 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 to create new question and 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 to students to try the system. 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 information can be seen in Fig. 7.

4.2.2.1. 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, he suggestions system provides help in selecting appropriate responses that already exist in the conversational model.

A screenshot of the interface is shown in Fig. 8. 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.

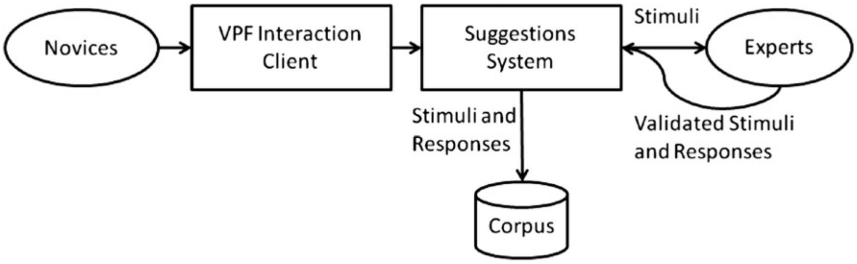
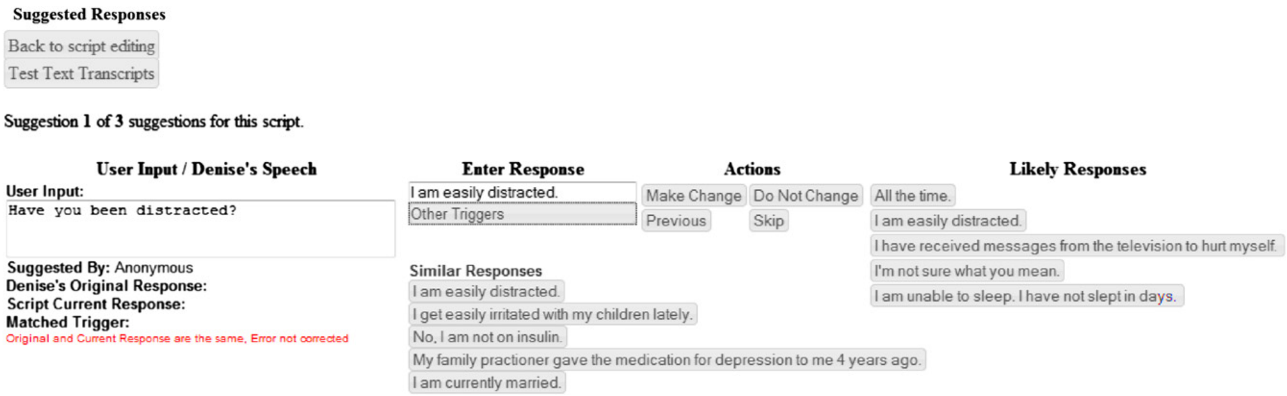


Fig. 7. Flow of knowledge in the VPF system.



Suggested Responses

Suggestion 1 of 3 suggestions for this script.

User Input / Denise's Speech

User Input:

`Have you been distracted?`

Suggested By: Anonymous
Denise's Original Response:
Script Current Response:
Matched Trigger:
Original and Current Response are the same, Error not corrected

Enter Response

I am easily distracted.
Other Triggers

Actions

Similar Responses

I am easily distracted.
I get easily irritated with my children lately.
No, I am not on insulin.
My family practioner gave the medication for depression to me 4 years ago.
I am currently married.

Likely Responses

All the time.
I am easily distracted.
I have received messages from the television to hurt myself.
I'm not sure what you mean.
I am unable to sleep. I have not slept in days.

Fig. 8. Suggested responses interface.

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.

4.3. *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 MADAL web-service (Rossen, 2010). In a standard MADAL web-service, each table of the database is represented as its own class. These classes use object relational mapping. This mapping allows clients of the web service to access any VH data in the database. 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, where 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 – each of these clients is described below.

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 Fig. 4, 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 account and script information, Second Life account information, and a location in Second Life for the VH to appear, 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. Currently, healthcare practitioners at the University of South Florida are using the *VPF*

Second Life application 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. 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. The Interpersonal Simulator has been used in several studies and has been validated for use in medical interpersonal skills education (Deladisma et al., 2007; Dickerson et al., 2005; Johnsen et al., 2007; Raij et al., 2007).

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.

5. 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 5.1)
- Evaluation 2: meta-evaluation of four published case studies that further examine student and educator real-world experiences with HDCM and VPF (section 5.2)
- Evaluation 3: usability evaluation based on self-reported feedback from experts on the efficacy of HDCM and VPF in real-world educational settings (section 5.3)

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 VH patients and educate students.

5.1. 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 and results in a conversational model with increased accuracy for spoken interactions.

5.1.1. 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 (previously shown in Fig. 3, 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.

5.1.1.1. Participants. *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 min and 25 questions.

5.1.1.2. Procedure. The pharmacy instructor uploaded a comma-separated list of student names and emails into VPF; 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 convenience.

5.1.1.3. 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 model is more accurate than the CCM model.

The HDCM 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 5.1.2, we show the number of unique questions and responses gathered 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 gathered 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 domain experts using interactions with 51 pharmacy students. Creation of the CCM model required six months and approximately 200 h. 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.

5.1.2. Results

5.1.2.1. Conversational modeling timer. Part 1 of the evaluation examined the progress of conversational modeling performed during the two weeks of development and how much 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 min, making the total student time 62 h. These three rounds of user-testing required 15 h of expert time (including 2 hs of training time and 13 h 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 1 alongside the results for the previously created CCM model.

5.1.2.2. 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. Fig. 9 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

Table 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 h (combined educator and knowledge engineers)	11 h (13 min average)	1418	303
HDCM	Pharmacy instructor & 186 pharmacy students	Virtual people factory: web-browser	15 h	62 h (20 min average)	2655	595

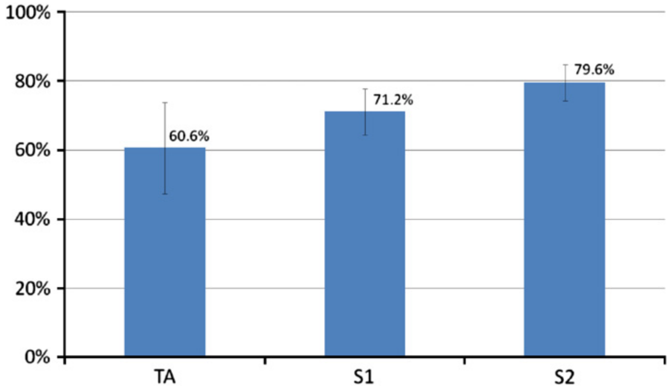


Fig. 9. The accuracy of the dyspepsia conversational model for each group TA s.d. = 13.3%, S1 s.d. = 6.7%, S2 s.d. = 5.3%, represented by the error bars above.

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 a previously created CCM conversational model. The final accuracy analysis with spoken inputs is provided in the next subsection.

5.1.2.3. 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 was able to transcribe spoken utterances at 83.3% accuracy. Transcripts of interactions from the previous study were processed for use in the current analysis.

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 Fig. 10.

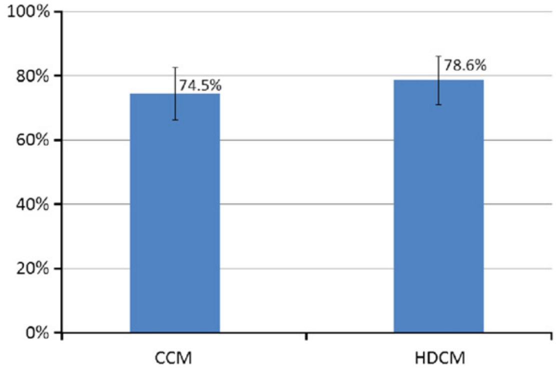


Fig. 10. Accuracy of the Centralized Conversational Model vs. Human-centered Distributed Conversational Model for 33 spoken transcripts, improvement of 4.1% is significant at $p < .05$. CCM s.d. = 11.1%, HDCM s.d. = 9.7% represented by the error bars above.

5.1.3. 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. We thereby see that these 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 h over two weeks, compared to the knowledge engineers and experts creating Vic in ~200 h over six months. Table 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

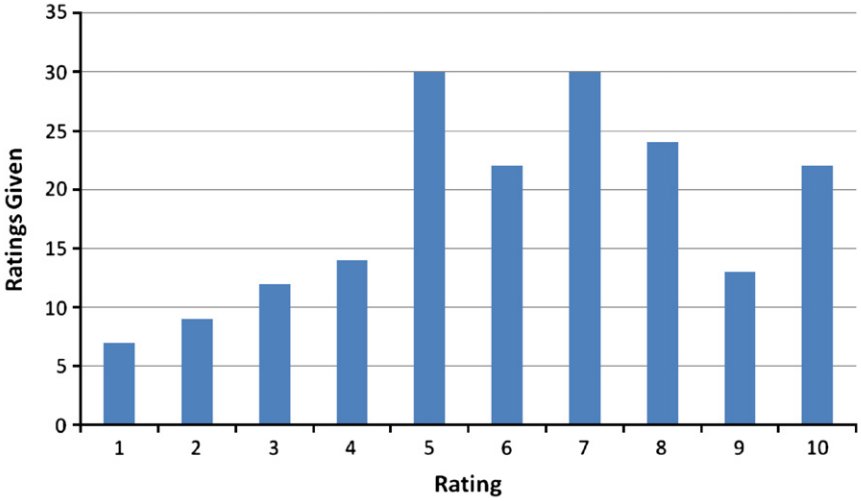


Fig. 11. Student participant ratings on the educational value of interacting with the virtual patient.

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 Fig. 11.

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."

5.2. *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 for teaching 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., 2010; Nantha Surkunalingam et al., 2009; Shah et al., 2012, 2009). Healthcare experts successfully created these conversational VHs with VPF with email only assistance from the VPF developers. 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.

5.2.1. *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 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 VH 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 as a supplement to traditional training methods such as standardized patient interactions and expert mentoring. These VH patient experiences provided additional opportunities for expert mentors to offer feedback to participating students.

5.2.1.1. Surgical patient with melanoma. This study evaluated VPF's perceived educational benefit for medical education. Researchers at the Georgia Health Sciences University evaluated VPF for use in their medical curriculum (Shah et al., 2009). 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.

5.2.1.2. Psychiatry patient with depression. This study evaluated VPF's perceived educational benefit for first, second, and third-year psychiatry students. Psychiatry researchers at the Georgia Health Sciences University 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.

5.2.1.3. 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 (Nantha 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. Fourty-sixsecond-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.

5.2.1.4. 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 Georgia Health Sciences University (Foster et al., 2010). 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 soon 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. Twenty five, 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.

5.2.2. Results

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

5.2.2.1. 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.

5.2.2.2. 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 4.

Open–ended feedback suggested that this VH 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.

5.2.2.3. 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 5.

Table 3
Percentage of participants reporting 7–10 (good-excellent) on 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%

Table 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 h
Depression patient	Cynthia Young	Psychiatry	1314	345	15 h
Bipolar disorder patient	Denise	Psychiatry	1605	220	11 h
Meningitis patient	Nelson Sanjaya	Osteopathy	777	228	25 h

Table 4
Percentage of participants reporting 4–5 (good-excellent) on 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%

Table 5
Percentage of participants reporting 4–5 (good-excellent) on 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%

Table 6
Comparison of content elicited during interactions.

	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%

5.2.2.4. 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 6. Participants interacting with the VH patient using a browser were more likely to ask about suicide, grandiosity, and elevated mood; while participants interacting with the VH 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.

5.2.3. Discussion

These studies indicate that VPF may be a viable and well-received method for augmenting current interview training curriculums. 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 three months to more than two-years full-time to develop a single virtual patient scenario, compare this to the 11–25 h of Table 2 to develop a VH 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 VH 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 world 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 VH 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 VH patients in the healthcare field.

5.3. 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

5.3.1. 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 is 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 are 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

Table 7
Results of the survey on educational value, usefulness, and ease of use.

	N	Mean	Standard 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
Clearn & 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

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.

5.3.2. Results

Seven out of the 11 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 7. Additionally, three out of seven domain experts responded that they will continue to use VPF in their courses (four responded that this option was not applicable as they were not teaching courses). Of the three domain experts who will continue to use VPF in teaching their courses, two out of three will use VPF to replace a portion of their lecture, and they note that this will save them on average 90 min of lecture time.

Further, to introduce a patient case, experts normally take 8–10 h to create the paper and pencil case. If we consider that a VPF VH patient takes 15 h to create, those five 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.

5.3.2.1. In discussing their interactions with VPF, healthcare experts made the following negative remarks:

Time consuming and accuracy issues

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

Limited animations and interactivity (in 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."

5.3.2.2. In discussing their interactions with VPF, healthcare experts made the following positive remarks:

Easy to use and enjoyable

- "[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."

Educationally valuable and recommended

- "[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."

5.3.3. Discussion

The results of evaluation three 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. We are currently working on further reducing the time required and improving the accuracy of interactions using conversation templates (see section 6). 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 (Raij et al., 2007). 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 the current 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 VH 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.

6. Limitations and future work

Thus far, we have identified two limitations to the HDCM approach: end-user availability and expert time requirements.

The main limitation of HDCM is that it 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 less educational value for students helping to create the conversational model.

A second limitation is that 15 h may still be a large amount of time for many experts. And while the time is shorter than the previously required 200 h, experts still perceive the process to be difficult.

A portion of the time to model a conversation goes into the initial seeding of the system. This seeding activity covers basic questions that are commonly used in conversations, such as "What is your name?" or "How old are you?". This costs time in both the seeding of the conversational model, and in the iterative refinement of the HDCM process. Many of the stimuli gathered in an initial round of testing are basic questions that are common to most medical interactions.

We are currently working on methods to reduce the time to seed the model and reduce the number of interactions required to model a conversation. One avenue of research is to reuse portions of previously modeled conversations to bootstrap the modeling of new conversations. With the current VPF system, the work that goes into creating one of these conversational models is not leveraged to create the next model. A future solution is to use conversational model templates. For example, we may use the set of questions from our Dyspepsia patient to create a *stomach pain template*. An expert can customize the template to create new responses to each stimulus, which will generate a patient for other stomach complaints (such as appendicitis, gallbladder infection, or ectopic pregnancy). These templates will ease the process of seeding the model and allow the first iteration of HDCM to gather stimuli that are more specific to the current conversation. These advantages may further reduce the time requirements for enumerating the stimulus-response space of a conversation, thereby increasing the accuracy of the conversational simulations.

7. Conclusion

The results of our study suggest that HDCM is both faster than CCM and can result in a more comprehensive enumeration of the conversational space. The techniques presented enable an expert to create a conversational model themselves, reduce conversational modeling time requirements, and result in a conversational model with increased accuracy for both typed and spoken interactions. HDCM speeds up the modeling process by enabling novice and expert users to create conversations with a VH in a distributed fashion. We found that this is an efficient method for generating VHs with the ability to recognize and respond to user speech.

We also see that VPF is an educationally valuable tool. Despite the time consuming nature of creating VH patients and the imperfect results of those VH patients, both students and healthcare educators report that the system is a viable method for educating healthcare students on their communication skills. Educators find that having control over the VH patient creation gives them the ability to provide focused learning experiences. And the educators who have used VPF in classroom education are continuing to use VPF in future classes.

Since these studies, our healthcare collaborators have continued using VPF to integrate VHs into the curricula of four medical schools in the United States. In August 2008, we opened VPF to the public: http://vpf.cise.ufl.edu. VPF currently has 56 active users outside of our research group, including VH researchers, healthcare practitioners, psychologists, and even high-school students. These users have explored creating characters outside of the healthcare domain including as educators and tour guides. From their work, as of September of 2011 VPF has facilitated over 2700 interactions consisting of more than 105,000 utterances.

Acknowledgments

Special thanks go to Dr. Carole Kimberlin and Dr. Diane Beck for their participation in the pharmacy study, and to Dr. Scott Lind, Dr. Adriana Foster, and Dr. Hevil Shah for their work on VH patients and their invaluable feedback on VPF. We also thank Dr. Aaron Kotranza, Dr. Regis Kopper, Dr. Andrew Raij, Dr. Kyle Johnsen, Joon Chauh, and Shiva Halan for their advice and assistance during VPF's development and the writing of this paper. This work was made possible by a University of Florida Alumni Fellowship and National Science Foundation Grants.

References

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

Bearman, M., Cesnik, B., Liddell, M., 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., Fors, U.G.H., 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.

Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13 3, 319–340.

Deladisma, A., Mack D., Bernard T., Oxendine C., Szlam S., Wagner P., Kruse E.J., Lok B. and Lind D.S. 2007. 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., Johnsen K., Raij A., Lok B., Hernandez J., Stevens A. and Lind D.S. (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.

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

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

Foster, Noseworthy, Shah, Lind, Lok, Chuah, Rossen, 2010. 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., Levin A., Westbrook C., Edwards-Daughtery P., Hurliman J., Taney, F., and Neal, J., 2010. Open Metaverse Foundation – libopenmetaverse from 〈http://www.openmetaverse.org/〉.

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

Hill, R., Gratch, J., Marsella, S., Rickel, J., Swartout, W., Traum, D., 2003. Virtual humans in the mission rehearsal exercise system. Künstliche Intelligenz 4 (03), 5–10.

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

Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., West, S.L., 2000. The virtual standardized patient. Simulated patient-practitioner dialog for patient interview training. Studies in Health Technology and Informatics 70, 133–138.

Iacobelli, F., Cassell, J., 2007. Ethnic identity and engagement in embodied conversational agents, Intelligent Virtual Agents (IVA). Springer, Paris, France.

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

Johnsen, K., 2008. Design and validation of a virtual human system for interpersonal skills education, Computer Information Science and Engineering. University of Florida. Doctor of Philosophy, Gainesville, FL 146.

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

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., Piepol, D., 2007. Building Interactive Virtual Humans for Training Environments. ITSEC, NTSA.

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

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

Lidwell, W., Holden, K., and Butler, J., 2003. Universal Principles of Design, Rockport.

Nantha Surkunalingam, J.W., Christopher Taranto, Samantha McCaskill, Christopher Blanchard, D., Scott Lind, Benjamin Lok, Brent 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.

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

Raij, A., Johnsen, K., Dickerson, R., Lok, B., Cohen, M., Stevens, A., Bernard, T., Oxendine, C., Wagner, P., Lind, D.S., 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.

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

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

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

Saleh, N., 2010. The value of virtual patients in medical education. Annals of Behavioral Science and Medical Education 16 (2), 29–31.

Shah, H., Rossen, B., Foster, A., 2012. Interactive virtual patient scenarios: an evolving tool in psychiatric education, Academic Psychiatry. Accepted, Awaiting Publication.

Shah, H., Rossen B., Lind, D., and Lok, B., 2009. 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.

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

Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., Li Zhu, W., 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.

Slater, M., Pertaub, D., Steed, A., 1999. Public speaking in virtual reality: facing an audience of avatars. IEEE Computer Graphics and Applications 19 (2), 6–9.

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

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

Villaume, W.A., Berger, B.A., Barker, B.N., 2006. Learning motivational interviewing: scripting a virtual patient. American Journal of Pharmaceutical Education 70, 2.

von Ahn, L. and Dabbish L. 2004. Labeling images with a computer game Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: 319–326.

Westberg, J., Jason, H., 2001. Fostering Reflection and Providing Feedback. Springer.

Yedidia, M.J., Lipkin, M., 2003. Effect of communications training on medical student performance. JAMA 290, 1157–1165.

Zanbaka, C.A., Ulinski, A.C., Goolkasian, P., and Hodges, L.F., 2007. Social responses to virtual humans: Implications for Future Interface Design Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: 1561–1570.