Synthetic Interviews: the Art of Creating a 'Dyad' Between Humans and Machine-based Characters

by
Donald Marinelli, Ph.D. and Scott Stevens, Ph.D.
Co-Directors, Entertainment Technology Center, Carnegie Mellon University

Synthetic Interviews is a technology developed at Carnegie Mellon University (CMU) in Pittsburgh, Pennsylvania, by Scott Stevens, Ph.D. and Michael Christel, Ph.D., computer researchers in CMU's School of Computer Science and Software Engineering Institute. Synthetic Interviews provide a means of conversing in-depth with an individual or character, permitting users to ask questions in a conversational manner (just as they would if they were interviewing the figure face-to-face), and receive relevant, pertinent answers to the questions asked. Existing Synthetic Interviews are accessible via either typed or spoken interfaces.

Through this exploration of the CG-persona, users are able to discover a character's behavior, likes and dislikes, values, qualities, influences, beliefs, or personal knowledge. The Synthetic Interview also strives to capture and convey the core human attributes of reflection, humor, perplexity, bewilderment, frustration, and enjoyment. Synthetic Interviews therefore attempt to create nothing less than a 'dyad.' A 'dyad' is any two individuals maintaining a socially significant relationship (though this is not to imply that Synthetic Interviews are to remain limited to one-on-one experiences; in fact, Synthetic Interviews utilizing multiple interviewers or interviewees are currently being developed).

To use the term 'dyad' (coined originally by the philosopher Martin Buber in his analysis of interpersonal relationships) is to state specifically that the Synthetic Interview aspires to establish a genuine, meaningful interaction between the user and the CG-persona (presuming, of course, that the Synthetic Interview itself is designed to be more than an enhanced 'Help' application existing solely to dispense information or engage in 'retail' transactions). Yet even here we need to redefine what we mean by meaningful. The higher human emotions are not the primary motivation. Rather, the striving is for a human/computer interface that allows for the conveyance of information complete with accompanying human cognitive and affective attributes. The shorthand for this is "lifelike."

Can we therefore create a human/computer interface via the Synthetic Interview that makes users believe that they are conversing with a fellow human? Pushing aside the fact that users will most likely be interviewing famous personages, Hollywood stars, entertainment celebrities, politicians, teachers, physicians, self-help gurus, or a deceased historical figure brought back to life via a total immersion performance (for example, Hal Holbrook performing his famous *Mark Twain* one-man show, Jason Robards as *Franklin*

D. Roosevelt, or Jerry Mayer becoming *Albert Einstein*), can a sense of openness, fun, trust and respect actually be created between user and CG-persona?

From the beginning, we have focused on achieving a lifelike Synthetic Interview by having an actual human figure respond directly to the user. We have captured human beings on videotape, then digitized, encoded, indexed, and otherwise processed them into our computer database, so that the Synthetic Interview users are indeed conversing with human figures, by virtue of their receiving recorded video clips of the human actor responding to the questions asked. These responses are presented in 'talking head' format. The challenge in doing this though has been to do it in a way that initiates, facilitates, sustains, and ultimately increases user involvement with the CG-persona.

[American readers will more easily grasp what we are doing if we describe our system as a really big "Jeopardy" game, i.e., a database of *answers* in search of the proper or appropriate *questions*. "Jeopardy" is one of the most famous American game shows, created by Merv Griffin back in the 1960s, and still running with great popularity throughout the country. It is completely text based.]

The challenge we have set for ourselves is turning this Synthetic Interview experience into one that may actually lead to a suspension of disbelief on the part of the user such that the user perception is that of an actual interview/conversation taking place. This "suspension of disbelief" is considered by many theatre practitioners to be the ultimate goal of performances done in the realistic (i.e., lifelike, naturalistic) style. This theatrical "suspension of disbelief" occurs in the theatre when an audience member brackets reality (i.e., the obvious fact that he/she is sitting in a theatre, and that everything on stage is make believe and preordained). Suspension of disbelief allows the audience member to experience empathy, sympathy, pity, joy, laughter, anger, and myriad other human emotions with the characters and story being acted out on stage.

The video production values that have been applied so far to Synthetic Interviews have reflected a lifelike look and aspiration. As mentioned earlier, the videotaped responses of the CG-personae have been in a "talking head" format. The actor playing Albert Einstein (Jerry Mayer) was videotaped sitting on a chair in a "blue screen" studio. A specific sitting position in the chair was selected from which the actor commenced and concluded all answers. In post-production a gradual fade in the image was applied to his torso, so that when presented on a 3-D projector Einstein would be perceived by users as floating in space, or some cosmic ether (which is what the image in fact appeared to be doing!).

The most significant video production work undertaken to date has involved efforts to maintain a sometimes contemplative, sometimes humorous, always busy CG-persona, who is active and alert between questions posed by the user, or between interviews. There are very few human beings for instance who sit perfectly still - ever! Achieving this living effect was done via standard post-production techniques, such as editing together scenes of the actor playing Albert Einstein thinking, jotting down notes, scratching his head, drinking from a cup of coffee, smiling at the user, laughing at a

private joke, or motioning the user to come closer. There was then a transitional sequence between these particular behaviors and the retrieved video clip that was in answer to the question asked. Within this "pool" of behaviors and actions we also included a variety of spoken Einstein aphorisms, i.e., various truisms and words-of-wisdom for which the great scientist is renowned. While successful to a degree, subsequent Synthetic Interviews will utilize explicit morphing technologies to achieve the same effect in a much more seamless manner.

Another important dynamic that allows a Synthetic Interview to approach the intimacy of a dyad is the fact that speech recognition establishes a more genuine 'first-person' style between the user and CG-persona. Speech recognition allows the user to become a 'protagonist,' someone whose decisions directly affect the course of the interview, and the nature of the relationship. The user does, to a large degree, control the direction and tone of the interview.

Using a spoken interface that allows users to speak with natural voice is essential in overcoming the interface barrier between user and CG-persona. Because of this demand, speech recognition software must be good enough to make the Synthetic Interview interaction seamless, as is the case when people are talking to each other. The Synthetic Interviews employ the Sphinx-II speech recognition system, one of the most reliable speaker independent, continuous speech recognizers in the world. Until recently, Sphinx-II could only run on a state-of-the-art workstation, but it has since been ported to a consumer Pentium PC running Windows 95.

The user poses questions by speaking into a microphone. The speech-recognition software, running on an ordinary Pentium PC of at least 90Mhz (120Mhz preferred) and 32 Megabytes of memory, analyzes the questions via its existing language models. We have found that for a typical Synthetic Interview a generic language model of 5,000 words and a domain specific language model of at least 1,000 words is usually sufficient. The generic language model is taken from the most common 5,000 English words and an analysis of their usage in everyday speech.

The domain specific language model includes idiosyncratic terms, proper names, foreign words and phrases, and geographic references that are part of the world of the persona being interviewed. For example, Einstein must respond to words like: relativity, atom bomb, violin, quantum, pacifism, sailing; and proper names such as Leonardo da Vinci, Galileo, Isaac Newton, Kepler, Adolf Hitler, Madame Curie, Nils Bohr, Nobel Prize; and geographic locales like: Ulm, Munich, Germany, United States, Italy, the Institute for Advance Studies at Princeton, the Bern Switzerland Patent Office. This gives highly accurate recognition of relevant questions while still giving acceptable results for out-of-bound questions. Taken together, speech and language understanding allows users to talk to the characters in a lifelike manner.

As stated previously, the Synthetic Interview structure that we have been experimenting with involves hundreds (potentially thousands) of pre-recorded video clips, i.e., a database of answers in search of questions. CG-persona responses may be

declarative or interrogative, but the key component to the creation of a lifelike dyad relies heavily on how one handles the problematic question, or when there is difficulty in understanding what precisely has been asked. To handle these instances, the Synthetic Interview must possess a sizable database of what we refer to as "default responses" and "pool responses." These are specific responses created and bundled together to handle such events as out-of-bounds questions, questions for which there are no answers, and/or non-interrogative statements that do not have a response.

The importance of "default" and "pool" responses cannot be underestimated. In fact, our studies have shown that it is precisely through these seemingly tangential actions that the suspension of disbelief is more readily achieved. A "default" response is a response that is triggered when there are speech recognition errors due to inadequate language models within the Sphinx II system. This pool of default responses include phrases like "Could you ask me that one more time?" "Could you please rephrase that question?", "I'm sorry, I didn't quite understand you. Could you ask it again?" or "Let's come back to that later?"

"Pool" responses are triggered when an "out-of-bounds" question is asked. The parameters of what constitutes "out-of-bounds" questions are determined and set during the Pre-Production and Production phase of Synthetic Interview creation. For example, in our celebrity demo it was clear that users might be inclined to ask a variety of obscene, lewd, or otherwise inappropriate questions. In anticipation of this event, the celebrity recorded various responses to these truly out-of-bounds questions. "Out-of-bounds" questions consequently are recognized, but are not answered directly. They trigger instead a response from a "pool" of video clips designed to shut down or stymie inappropriate questions or statements. Both the "default" and "pool" indices have between 15 and 30 responses assigned to each, the aim being to cut down on repetition.

We are enhancing the lifelike quality of our Synthetic Interviews by creating various transitional phrases that can change invalid statements to valid ones, such as the phrase, "I don't really know about that, but let me discuss something else of interest." Or "Let's come back to that, but in the meantime why don't we chat about the following."

The degree to which the CG-persona is able to be pro-active (i.e., being able to guide the interview in certain directions) is also of utmost importance. The fact is that most spoken human communication (outside of a classroom at least) is turn-taking, a give-and-take wherein one person speaks in either a declarative or interrogative manner about a specific domain or topic, and then the other person acts out the very same sequence of events until closure on the subject matter or the communication in general is initiated by one or both parties. The very nature of the Synthetic Interview though puts the onus for interaction (at least initially) on the shoulders of the user, i.e., the person asking the questions. To create a more lifelike interface however, requires us to endow the CG-persona with both commentary possibilities, and the ability to initiative his/her own questioning.

Of the Synthetic Interview demonstrations we have created so far, only our celebrity interview has pro-active capabilities. In this instance, the celebrity can be programmed to suggest an area of discourse to the user after a certain time period of no activity. For example, if 30 seconds passes after the user last asked a question, the system will trigger the celebrity to ask one of the following statements/questions: "Why don't you ask me about my education?" or "Why don't you ask me about acting?" or "Why don't you ask me about my family?" The pro-active question is chosen randomly from a pool of pro-active statements. In the newest Synthetic Interview under development, prior discussion within a certain domain will prevent that particular pro-active statement from being triggered. In other words, prior discussions about the actress' youth and training in acting will prevent the pro-active statement, "Why don't you ask me about my education" from being triggered.

Ultimately though, the one factor contributing the most to a truly lifelike quality within the Synthetic Interview, thereby enhancing the possibility of a user/CG-persona dyad being created, is in the performance and acting qualities of the recorded answer. In this respect it all comes down to the traditional acting and directing skills. In other words, the more lifelike the videotaped response, the more engaging the Synthetic Interview interaction. Traditional psychological research has revealed on numerous occasions that as much information is conveyed subliminally or via visual cues as is conveyed via speech. For example, a slight pause before answering a particularly "tough" question (for instance, one involving religious, philosophic, personal, or otherwise challenging issues) is a visual cue indicating thought, reflection, consideration, and deliberation of the question asked. The corresponding reaction in the user is usually one of empathy, an endearing human attribute that can make what had begun as a simple Synthetic Interview transcend into a more personal meaningful interaction. We are therefore on the road towards creation of a dyad.

What then are the potential uses of Synthetic Interviews? It is our belief that Synthetic Interviews have direct applications to all aspects of interactive entertainment. The first Synthetic Interview products will seek to capitalize on the celebrity phenomenon that is so much a part of American culture. Talk shows, call-ins, *PEOPLE* and *US* magazine, and other celebrity venues in every aspect of our culture underscore the desire on the part of so many American demographic groups to have personal access to a movie or television star, singer, sports figure, teacher, politician, religious leader, self-help guru, physician, etc. Any person possessing a large enough fan base of potential users could profit from creation of a Synthetic Interview, either as a stand alone CD-ROM, World Wide Web site, or self-standing kiosk.

The ability to achieve a dyadic effect through a Synthetic Interview has even greater importance for those instances when the CG-persona is neither a public celebrity nor a historical figure, but is instead a family member. One aim is to create Synthetic Interviews that capture for posterity a parent, grandparent, or great-grandparent recollecting life experiences, remembrances, advice and wisdom, in a way that allows the inner affective life of that individual to shine forth. Being able to create a Synthetic Interview that facilitates and establishes a 'dyad' between deceased family members and

their never-seen progeny would be an important triumph for technology in the service of humanity.

Donald Marinelli & Scott Stevens
Pittsburgh, Pennsylvania
January 1998