Lessons Learned in Modeling Schizophrenic and Depressed Responsive Virtual Humans for Training

Robert C. Hubal rhubal@rti.org
Geoffrey A. Frank gaf@rti.org
Curry I. Guinn cig@rti.org
Technology Assisted Learning Division · RTI
3040 Cornwallis Road
Research Triangle Park, NC 27709

ABSTRACT
This paper describes lessons learned in developing the linguistic, cognitive, emotional, and gestural models underlying virtual human behavior in a training application designed to train civilian police officers how to recognize gestures and verbal cues indicating different forms of mental illness and how to verbally interact with the mentally ill. Schizophrenia, paranoia, and depression were all modeled for the application. For linguistics, the application has quite complex language grammars that captured a range of syntactic structures and semantic categories. For cognition, there is a great deal of augmentation to a plan-based transition network needed to model the virtual human's knowledge. For emotions and gestures, virtual human behavior is based on expert-validated mapping tables specific to each mental illness. The paper presents five areas demanding continued research to improve virtual human behavior for use in training applications.

Categories and Subject Descriptors
D.2.1 [Software Engineering]: Requirements/Specifications---*elicitation methods, methodologies*; E.1 [Data Structures]---*graphs and networks, tables*; H.5.2 [Information Interfaces and Presentation]: User Interfaces---*natural language*; J.4 [Computer Applications]---*psychology*.

General Terms
Documentation, design, experimentation, languages.

Keywords
Agents, behavior modeling, managing encounters with the mentally ill, interaction skills training, responsive virtual humans.

INTRODUCTION
With federal and internal funding, RTI developed a prototype training application called JUST-TALK to provide law enforcement personnel practice in conversing with the mentally ill. We created five basic scenarios. In all scenarios the officer receives a report of a young, adult male who is reported to be behaving erratically, entering the street, and almost getting hit by a car. The officer encounters the virtual human on a city street; the virtual human may be sitting on a bench, standing in front of or behind it, or pacing around it. Through a natural language dialog with the virtual human, the officer is supposed to infer whether the virtual human is schizophrenic, paranoid, depressed, sad but otherwise normal, or stressed but otherwise normal. Depending on the officer's diagnosis, the virtual human may be released or asked to step into the patrol car to receive some help. If the dialog goes poorly, the virtual human may run away, attack the officer, or enter into a catatonic state.

JUST-TALK was first fielded during a course titled "Managing Encounters with the Mentally Ill" in January 2002 at the North Carolina Justice Academy (NCJA) [12]. The NCJA was a logical venue for testing the application and technology, since it has the mission to improve the quality and effectiveness of North Carolina's criminal justice services through research, education, and training.

JUST-TALK is one of a series of responsive virtual human applications RTI is developing for interaction skills training. We and our customers feel that lifelike agents are likely lead to competency or mastery of subject matter; we reason that the realism of interacting with an emotive, responsive virtual human will engage the student and will lead to effective acquisition and greater retention [20,21]. We argue, furthermore, that using responsive virtual humans in conjunction with classroom-based training and student role-plays represents a cost-effective training approach [19].

In this paper we describe lessons we have learned from modeling behavior not just of a normal individual, but also schizophrenic, paranoid, and depressed behavior. We discuss linguistic, cognitive, emotional, and gestural issues.

AGENT SOFTWARE
Practicing skills in a safe and supportive environment allows the student to learn flexible approaches. Flexibility is critical for interaction skills [14] and for performing operations and maintenance under difficult conditions, such as time constraints, dangerous, and information-poor conditions [24]. The consistency that is gained by repeating this practice in virtual and constructive environments leads directly to good decisions on the job [24,37]. Practice also leads to increased confidence prior to the first real on-the-job experience.

We have developed a PC-based architecture, Avatalk, that enables users to engage in unscripted conversations with virtual humans and to see and hear their realistic responses [20,21]. Among the components that underlie Avatalk are a Language Processor, a Behavior Engine, and a Visualization Engine (Figure 1). The Language Processor accepts spoken input and maps this input to an underlying semantic representation, and then functions as a speech generator by working in reverse, mapping semantic representations to speech output, facial expressions, and gestures, displayed by the Visualization Engine. The Behavior Engine maps Language Processor output and other environmental stimuli to agent behaviors. These behaviors include decision making and problem solving, performing actions in the virtual world, changes in facial and body expression (via the Visualization Engine), and spoken

IUI'03, January 12-15, 2003, Miami, Florida, USA.

dialog. The Behavior Engine also controls the dynamic loading of contexts and knowledge for use by the Language Processor. The Visualization Engine takes gesture, movement, and speech output and enables the 3D representation of a human to perform these actions. It accomplishes these movements through morphing of vertices of a 3D model and playing of key-framed animation files (largely based on motion capture data). The Visualization Engine is capable of lip-synching to both synthesized and recorded speech.

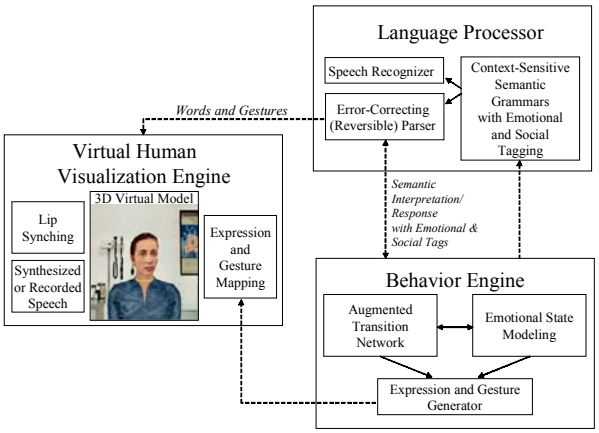


Figure 1: Avatalk Architecture

Language Processor

The Language Processor maps spoken input to underlying semantics and produces responsive output. It has six components [17]:

- Speech Recognition.
- Parsing. Our system uses a minimum distance translator parser that tries to match the spoken words to the closest grammatical sentence as defined by the currently active language grammar [16].
- Dynamic Grammars. A grammar specifies the acceptable spoken statements. The representation language is quite free; literally any sentence can be encoded in the grammar. In Figure 2, the sentence "help me find the spaceship" will return the semantic statement "`ask(location(ship))`". We use semantic categories to categorize syntactic components, an efficient parsing strategy that greatly assists in handling ambiguity with the main disadvantage of being domain specific. The Behavior Engine dynamically selects which grammars should be active based on the current context.
- Language Interpretation. The considerable ambiguity inherent in language, and reliance on linguistic context, makes natural language processing difficult. For instance, taken alone "it is" has very little meaning, but following a command "put the knife down" it makes sense as "the knife is on the ground". In an environment where there are multiple knives "the knife" is also ambiguous. We resolve these ambiguities via dynamic context switching and use of utterance expectations sorted by likelihood [41].

```
S -> ASK LOC' : ask(location(LOC')) .
ASK -> PLEASE WHEREIS .
ASK -> WHEREIS .
PLEASE -> damn it @-0.4 .
PLEASE -> please @ 0.2 .
PLEASE -> would you please @ 0.3 .
WHEREIS -> help me FIND .
WHEREIS -> where is .
FIND -> find .
FIND -> locate .
LOC -> the SPACE ship : ship .
LOC -> the rest room : wc .
SPACE -> mother .
SPACE -> space .
```

Figure 2: Sample Language Grammar

- Reliability Scores. We compute the likelihood of having correctly understood a particular utterance, factoring in the speech recognizer score (a "goodness of fit" between the audio signal and the acoustic model it has of spoken language), parser score (a minimum distance parser returns a score based on the number of insertions and deletions needed to parse the string of words using the language grammar), expectation score (based on context), goal weight (certain critical goals may require exacting verification), and past recognition rate. Depending on a dynamic threshold, the system will ask the user to repeat himself/herself, paraphrase what it believes was said and ask for confirmation, or accept its interpretation and continue.
- Emotional and Social Tagging. We extended the semantic grammars by applying tags that carry information related to emotional and social Behavior Engine state variables. In Figure 2, the symbol '`@`' is used to indicate the relative `POLITENESS` of different phrases. (We have also mapped `CONFUSION`, `SATISFACTION`, `HUMOR`, time constraint, and other tags in our applications.) Values range from -1.0 (very impolite) to 1.0 (very polite); if not specified, the word or phrase has a neutral value related to that attribute. During parsing, these values are combined to produce a final score.

Behavior Engine

The Behavior Engine uses semantic interpretation generated by the Language Processor to assist in determining virtual human behavior. The current underlying architecture of the Behavior Engine is an augmented transition network (ATN). Typical of ATN, there are often multiple conditional transitions leading between network nodes; at least one transition condition defaults to true. If multiple transition conditions are satisfied at a particular node, then one is selected at random.

One important set of variables that we maintain in the ATN are cognitive, generally domain-specific, variables. These variables are used while tracking conversational topics and in the interface with the Language Processor to expect or generate relevant statements. Another set tracks physical or physiological characteristics; in certain applications we use physiological models that provide continuous, real-time cardiovascular, respiratory, and pharmacological simulation [23]. The virtual human can exhibit real-time medical signs and symptoms, which can in turn affect other behaviors.

Another set is emotion variables. We keep track of a base set of emotions and personality traits for each virtual human in the simulation. Combinations of values from this base set are used to define all emotional state descriptions. For instance, a base set that we use is `ANGER`, `CONFUSION`, `DISGUST`, `FEAR`, `HAPPINESS`, `HUMOR`, `POLITENESS`, `SADNESS`, `SURPRISE`, `TIME_CRUNCH`, and `VOLATILITY`. We then define other emotions based on these (as well as constants such as `SMALL`, `MEDIUM`, and `LARGE`; Figure 3a), and emotional states iteratively based on emotions and other emotional states, using Boolean expressions (Figure 3b); these emotional state descriptions are used in node condition statements. In node action statements, we allow modification of emotional state (Figure 3c). For instance, if a node is visited only when the user commits an error, or when the virtual human is distracted by an action in the virtual environment, the value of emotion values may change. Thus, the emotional state of each virtual human is dynamic and depends on current state, environmental constraints, and user performance.

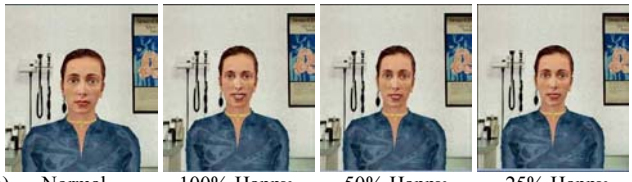
The choice of base set is somewhat arbitrary, though we used as its core a commonly accepted set [11,34,40]. The equations used to define emotional states derived where possible from the research [10,39], otherwise from expert advice, common intuition, application demands, and ad-hoc experimentation.

a)
```
DISTRUST -> (SIZEABLE*(ANGER+FEAR))-(MEDIUM*SURPRISE) .
SATISFACTION -> (VERY_LARGE*HAPPINESS)-(MEDIUM*ANGER) .
```
b)
```
CERTAIN -> DISTRUST <= EXTREMELY_LOW .
COMPLACENT -> SATISFIED or not(CONFUSED) .
HOSTILE -> (ANGRY and DISGUSTED) or not(FRIENDLY) .
```
c)
```
IRK -> ANGER += VOLATILITY*TRIVIAL .
PRESSURE -> TIME_CRUNCH += VOLATILITY*LARGE .
```

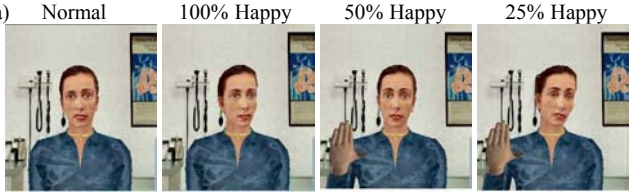
Figure 3: Emotional State Computations

Visualization Engine

The Visualization Engine produces our 3D virtual humans. Target images created for the extremes of emotional expressions, gestures, phonemic facial movements, and other body movements feed into our own morphing algorithms; we create real-time movements by morphing from the original image to these morph targets. For instance, if we create a HAPPY morph target, we can generate a face that is completely happy or at any level between the original image and the target (Figure 4a). Further, we can blend any morph targets together, so we can combine a 50% SARCASTIC face with a 75% DISTRUSTFUL face with a 90% HAND_WAVE (Figure 4b).



a) Normal 100% Happy 50% Happy 25% Happy



b) 50% Sarcastic 75% Distrustful 90% Hand Wave All 3 Combined

Figure 4: Happy and Combined Facial Expressions

The Behavior Engine produces emotional and gesture expressions based on the state of the simulation. For instance, the virtual human may be at a distance salutation node, where it is instructed to wave [8]. These commands are sent directly to the Visualization Engine. An additional mechanism within the Visualization Engine automatically generates appropriate facial expressions and gestures based on the virtual human's emotional state, and an analysis of textual structure of the virtual human's utterances. Thus an ANGRY virtual human may produce an angry expression even if the Behavior Engine does not explicitly specify this action. Similarly, positive and negative responses will invoke head nods and shakes or other emblems [18], predefined terms will invoke propositional gestures to complement the utterance [5], beat gestures are inserted [5], and eye movements that reflect gazing are modeled after prototypical human behavior [8]. Finally, both the Behavior and Visualization Engines insert some small random head and body movements to add realism.

Our architecture allows us to support both synthesized and recorded speech. We employ off-the-shelf text-to-speech synthesizers, using to the extent possible the virtual human's emotional state to change synthesis parameters such as rate of speech and volume. A speech synthesizer has great advantages in flexibility of design; unfortunately, available speech synthesizers lack the fidelity of real human speech. Initial impressions of users tend to be negative and distracting. Only with continued use do users tend to adapt to the stilted speech (until there is a blatant mispronunciation). Therefore, with applications where the user has only short interactions with the system, we use recorded voice instead of synthesized voice. Rather than try to piece together fragments of recorded sentences, we generally record entire utterances. With recorded speech, the grammars used for speech generation have pointers to the associated sound files, as well as gestural commands. While this method is much less flexible than synthesized speech, the resulting simulation fidelity is greatly increased.

JUST-TALK Application

JUST-TALK teaches students basic techniques for managing encounters with the mentally ill by having them work through a series of one-on-one scenarios with a simulated subject [12]. It also teaches them to look for indications of particular forms of mental illness so that they can adapt their responses appropriately. Through observations of the virtual environment and a dialog with the virtual subject, the student must stabilize the situation and decide whether to release or detain the subject.

The JUST-TALK virtual environment is the sidewalk in front of a hardware store, where there is a bench. A patrol car is parked on the side street next to the hardware store. The subject is a white male adult. JUST-TALK has been implemented with five scenarios:

- A schizophrenic who is hearing voices;
- A paranoid who fears the police are conspiring with federal agents;
- A normal individual who is agitated from nearly being run over;
- A depressed individual who has become suicidal because of marital and child custody problems;
- A normal individual who is depressed in trying to deal with marital and child custody problems.

A session with JUST-TALK starts with the dispatch "You are responding to a report of a young, adult male who is reported to be behaving erratically, entering the street, and almost getting hit by a car.". The virtual environment for the scene is displayed in the scenario window. First, the user is expected to introduce himself or herself to the subject, although a novice may skip this step and start asking questions or demanding responses.

After the introduction, the user can then interview the subject. The interactions with the subject in JUST-TALK are all verbal; it does not teach apprehending or even officer safety techniques. Either the subject or the user may initiate the dialog. Sometimes the subject may be very withdrawn, so the user will have to open the conversation. Other times the subject may be very agitated and will start talking at the user from the start. The user uses the conversation to stabilize the situation, assess if the subject is anchored in reality, and determine what action is needed (e.g., leave, transport the subject to a mental health facility, or subdue the subject).

The user can stabilize the situation by talking with the subject to determine the problem and getting the subject to agree to a solution (i.e., asking the subject if he or she is taking prescription drugs, asking if he or she hasn't taken a dose recently, then persuading the subject to reinitiate his or her medication, or else visit a mental health facility), or talking with the subject and acknowledging that the subject's delusions are real to the subject and offering to help (but not agreeing with the delusions). The user can destabilize the situation by using inflammatory language or by challenging delusional or hallucinatory statements. Authoritative, commanding language can actually escalate the intensity of the interaction, particularly with subjects who are paranoid, distrustful, or afraid. Language that is more conciliatory, such as expressions of understanding, or requests rather than commands, can result in reduced tension. The user can assess if a subject is delusional by asking the subject about his or her mental illness history, listening for delusional or hallucinatory statements, seeing if the subject can respond rationally to questions about the problem and the subject's physical status. Users should also make note of physical gestures such as head movements, eye movements, and other body

language. Often, a subject who is hearing voices and other sounds (or taking some anti-psychotic medications with visible side effects) will display distinct physical signals.

INITIAL BEHAVIORAL MODELS

For the JUST-TALK application, we set out to create a virtual human that acts as if he is schizophrenic, paranoid, depressed, sad, or stressed. Though we have used them in the past, this virtual human was not intended to be a pedagogical agent [22,28]. Instead, the encounter was to take place, in a virtual environment, under the premise of interacting in a public space with a law officer. To achieve such a virtual human required that we first understand how normal (calm, aware, attentive) individuals would behave in a similar situation, then modify normal behaviors as appropriate for the scenario.

Basis in Agent Research

Some three-quarters of all verbal utterances are accompanied by gestures [6]. In addition, eye gazing and body posture play crucial roles during interactions. Our Behavior Engine attempts to mimic these activities to make the virtual human appear realistic.

Of the gestures, the majority are representational (iconic, deictic, emblematic, or metaphoric) and the rest are beats and idle motions [6]. Iconic movements are meant to convey information about spatial relationships or concepts [5]. Deictic movements, like pointing, are used mainly when discussing a shared task [5]. Emblematic gestures are culturally specific (such as a nod meaning "yes" or thumbs-up for "good") [18]. Metaphoric gestures commonly accompany new segments in communicative acts, and thus, like most representational gestures, rely on semantic knowledge [5,6]. In fact, representational gestures are often begun before the utterance even begins, as soon as the speaker knows what s/he is going to say [6]. Beat gestures, on the contrary, rely on syntax and prosody, occurring with heavily emphasized words or on occasions of turning over the floor to another speaker, though they may also convey information about novelty in discourse [5]. Idle motions (habitual actions such as winding or checking one's watch, lighting a cigarette, putting hands in the pockets, or manipulators such as stretching, wetting the lips, and scratching the head) are randomly executed throughout the interaction [18].

Eye gazing helps regulate the flow of the conversation. Looking straight at the conversational partner after an utterance implies seeking feedback [5], while staring is meant to intimidate [6]. Averted gaze can indicate sadness, depression, embarrassment, or confusion [6]. Blink rates change based on emotion, so that the normal blink per four seconds increases to one per two seconds when the individual is nervous, but decreases to one per six seconds when angry [31]. Together with facial expression, eye gazing can provide meaning to initiation and termination of a conversation, turn taking, and feedback [5]. For instance, inviting contact involves a sustained glance and a smile, while breaking away involves glancing around [8]. Similarly, an introduction often includes tilting the head, giving a turn includes looking at the partner and raising the eyebrows, wanting a turn includes raising the hands into view and looking at the partner, and planning a response involves looking away and lowering the eyebrows [8,9]. Positive feedback often involves nodding, while negative feedback may involve gazing away and increased blinking rate [9].

The degree of eye opening, position of eyelids relative to the irises, position and shape of eyebrows (arched, raised, drawn together), and other facial movements can be used to indicate emotion [26]. For instance, surprise is shown with wide open eyes, the lower eyelids drawn down, raised eyebrows, and the mouth open wide [9,31]. Similarly, fear is shown as wide open eyes and mouth, the upper eyelids raised, the lower lids tensed, and perhaps a step back or skin paling or sweating [9]. Happiness is shown as wrinkles below the lower eyelids, the lids raised but not tensed, smiling, the head lifted, and open body orientation [6,18], whereas contempt is shown as a sneer, a wrinkled nose or wrinkles under the eyes, the upper eyelids partially closed, and the body turned to the side [9].

The precise position of the body or one of its parts (i.e., posture), compared to a determined system of references, holds great meaning. For instance, the bodily attitude of prostration with head bent and shoulders falling is typical of unease [18]. Postural positions that have been well described include attentive, relaxed, insecure, confused, angry, joyful, mocking, insulting, rejecting, and welcoming [18].

Visualization based on Games

We quickly came to the realization that our rendering engine was not powerful nor flexible enough to satisfy all requirements for a normal virtual human, much less a mentally ill virtual human. Realistic gestures, it turns out, are at least if not more important during a conversation than simply emotional expressions [7,9]. For instance, in a virtual asthma patient application [21], whenever the virtual human delineated a list of items (e.g., "Colds or flu, exercise, perfumes, hair spray, pollens from grass and weeds, and house dust all bother me.") we wanted her to tap her hands or count with fingers, but the modeling effort required to allow her to delineate with gesture was prohibitive. Similarly, in a bank teller training application [demonstrated at 15], we wanted to customer to hand over identification and a check, but settled for a magical appearance of these items when requested.

Our Visualization Engine lacked these capabilities because it was designed primarily for facial expression. When we began developing Avatalk, we envisioned basically a talking head with which the user interacted, and there were no products capable of portraying the range of emotion we felt was desirable. We invested in our own rendering engine. The demand from users, though, was for a much more complete environment; we could not achieve the level of engagement we needed from only a talking head. So, for instance, we were told by JUST-TALK subject matter experts that the schizophrenic person should pace and point and move constantly, while the depressed person should sit and rock when he becomes agitated.

REVISED BEHAVIORAL MODELS

Though we discuss the linguistic, cognitive, emotional, and gestural models separately, in fact they all interact. This knowledge is not new; other researchers have found that emotions interact with social, perceptual, motivational, and motor systems [10,25].

To spell out the interactions, we developed, with some expert help, a series of tables that guide virtual human behavior. Table 1a shows a sample table for emotional state changes based on user input. Table 1b shows a sample table for determination of reply type based on current emotional state. Table 1c shows a sample table for gestures, also based on current emotional state. We implemented the tables in the ATN as condition/action statements. Our action definitions and definitions of emotional state changes in the ATN, and determination of next behavior(s), are very similar to the implementation described in [38].

Linguistic Models

We extended our linguistic analyses considerably in the JUST-TALK application. For user input, we built quite complex language grammars that captured a range of syntactic structures

and semantic categories. For virtual human output, we devised an extensible method of labeling phrases that increased productivity, complexity, and capability for reuse.

Table 1: Mapping Tables

a) Emotional state transitions, Depressed individual, based on user input

Current State of DEPRESSION	Next State for DEPRESSION	Drivers
SUICIDAL	DEPRESSED	Personal Request, Inform Help, Statement of Concern
SUICIDAL	DISCOURAGED	Threat, Insult, Profanity
DEPRESSED	DISCOURAGED	Personal Request or Query, Inform Help, Statement of Concern, Threat, Insult, Profanity
DEPRESSED	SUICIDAL	Command, Impersonal Query
DISCOURAGED	STABLE	Personal Request, Inform Help, Statement of Concern
DISCOURAGED	DEPRESSED	Command, Impersonal Query, Threat
STABLE	DISCOURAGED	Command, Impersonal Query, Threat

b) Reply mode map, Schizophrenic, for Anger & Fear

	APPEASED	TICKED	ANGRY	ENRAGED
TERRIFIED	Deny	Deny	Question	Challenge
AFRAID	Deny	Deny	Question	Challenge
SCARED	Respond	Respond	Challenge	Challenge
CALM	Respond	Respond	Challenge	Challenge

c) Gesture map, Paranoid, Question replies, for Anger & Fear

	APPEASED	TICKED	ANGRY	ENRAGED
TERRIFIED	n/a	Run away	Run away	Run away
AFRAID	Lean forward Look down Brace arms Stand behind bench	Lean forward Look forward Brace arms Stand behind bench	Torso upright Tilt head Arms at sides Pace	Get ready to fight
SCARED	Torso upright Look down Clasp hands Stand before bench, ready to comply	Torso rocks Tilt head Clasp hands Stand before bench	Torso upright Tilt head Cross arms Pace	Get ready to fight
CALM	Torso upright Tilt head Clasp hands Sit, ready to comply	Torso upright Tilt head Cross arms Stand before bench, ready to comply	Torso upright Tilt head Hands on hips Pace	n/a

As we (and others [32]) have for other applications, we decomposed the interaction into an introduction, interview, and resolution. We have found that, for a given domain, the introduction is rather formalized. For instance, in a clinical setting [21,23], the practitioner will usually begin with the standard "Hello, how are you?", whereas in a formal field interview setting [14], the interviewer is taught to begin with a scripted phrase ("Hello, my name is…, I represent RTI, a non-profit research institute located in North Carolina, I am in your neighborhood conducting a survey sponsored by…, you should have received a letter about it."). In JUST-TALK, the officer is expected to introduce himself/herself as he/she would to the man on the street.

During the interview, the officer is expected to de-escalate the potentially explosive situation, by using calming, polite, responsive language. Given appropriate user input, the virtual human will become more calm and composed and responsive, perhaps demonstrating this by stopping pacing or sitting down. Topics that the officer and subject will generally discuss include the event leading to the dispatch, the subject's name, family, illness (if any), and medications.

In other applications, the resolution is normally a simple "goodbye" or "thank you". In consultation with our subject-matter experts, we settled on five possible resolutions: no action is taken (the officer says "goodbye"), the subject runs away, the subject verbally attacks the officer, the subject descends into a catatonic state, or the subject is persuaded to get in the patrol car to be taken to a mental health facility for diagnosis, observation, or treatment.

Avatalk applications use spoken natural language interaction [30], not text-based interaction [29]. We observed a considerable number of preliminary users and collected their actual phraseology. Grammar definition became an iterative process, with each redefinition subjected to expert input to assess accuracy, relevance, and comprehensiveness. Still, we have work before us to increase recognition of user input, and to take further advantage of requests for repetition or confirmation.

User Input

We created a grammar framework whereby the JUST-TALK Language Processor could, mainly by examining syntactic structures, infer what type of input the user uttered. Different types included commands, requests for information, statements of understanding or appreciation, and informative declarations. From this syntactic analysis, we calculated a user politeness score and an input complexity value. Note that these scores are somewhat domain specific. For instance, our experts informed us that law officers are taught to always use "sir" or "ma'am" when conversing with all adults, hence any input lacking these terms detracts from the politeness count. In a field interviewing domain, on the other hand, any input including these terms would increase the politeness score. Similarly, how complex an utterance is depends on other phrases in the domain [see 35].

In the Behavior Engine, we then used contextual knowledge to evaluate the semantics returned by the Language Processor. From this semantic analysis, we derived a relevance metric and a personalization score. The relevance metric tells how appropriate was the user input based on the current topic (determined from previous input and from the most recent virtual human output). The personalization score provides an estimate of how well the user tailored his/her response to the virtual human; this measure, too, is domain-specific.

In general, we've found that a reply or response can be described by numerous adjectives, some of which are listed in Table 2. The descriptors listed in the left column are those that we are able to derive from syntactic and semantic analyses. (The descriptors listed in the right column represent ongoing research, requiring technology beyond what is readily available or implemented.)

Table 2: User Input Descriptors

Accurate	Childlike
Complete	Emotional
Deceptive	Feminine/masculine
Expected	Hesitating
Humorous	Instantaneous
Linguistically complex	Loud
Misunderstood	Non-native
Personal	Nonverbal
Polite	Out of breath
Positive	Sarcastic
Relevant	Sick
Ungrammatical	Tired
Verbose	Untruthful

Virtual Human Output

We structured the dialog so that the virtual human replies in one of six ways: response, question, denial, challenge, show of confusion, or zone out (see Figure 5). We used fuzzy logic in the Behavior Engine to decide on reply format, and used information on semantic content of the input from the Language Processor to select a specific appropriate verbal reply.

In JUST-TALK, we tagged the following emotions to the grammars: ANGER, ANXIETY, ATTENTION, DEPRESSION, FEAR, and HOSTILITY. This base set, derived through some discussions with experts, seemed to capture the range of emotional states and personalities we needed to portray. Work remaining to do, though, includes making some virtual human replies less coherent and more abusive, as appropriate for its personality, and integrating recorded speech as opposed to text-to-speech generation, to increase realism and user engagement.

Cognitive Models

We keep track of domain knowledge in the ATN via state variable settings, and also by its very structure, since some level of planning is inherent in the ATN (as opposed to using a modeling language [13]). Our virtual humans reason about social roles and conventions [36] through the ATN structure (what can be stated or asked at any point in the dialog) and grammar definitions (how it gets stated or asked). Figure 5 shows how we map language input to different sections of the ATN.

The architecture was designed to allow the application creators flexibility in assigning general and domain-specific knowledge. For instance, our virtual humans may not understand what "SATISFACTION" or "EXTREMELY_LARGE" mean, but they behave as if they do. Similarly, our virtual asthma patient [21] discusses relevant symptoms based on a specific setup variable indicating severity level, while the JUST-TALK subject portrays paranoia about the federal government or distrust of law enforcement only in relevant scenarios, even though the ATN structure is identical for all scenarios.

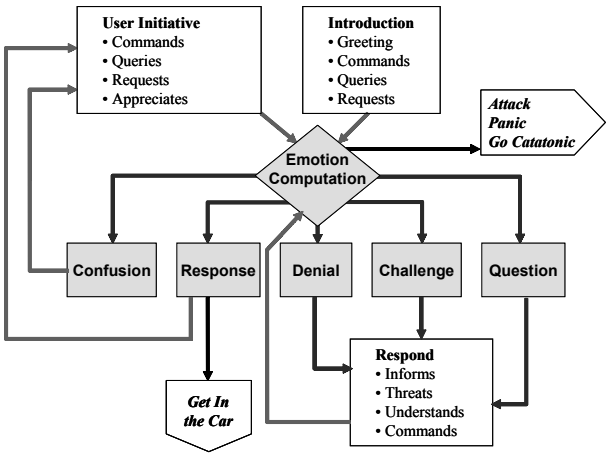


Figure 5: Grammar to ATN Mapping

Emotional Models

Our emotion models were built using several emotion and personality theories, including the Five Factor Model [2,42], Circumplex theory [34], and cognitive theory of emotions [33]. The latter model underscores most of our work, providing a scheme for labeling common emotions based on how our virtual humans react to inputs, events, and objects [1]. We also include an "emotion reasoning architecture" to describe how personalities can change over the course of an interaction [10], though we refrain from rapid large fluctuations in emotional states since users usually believe in virtual humans who have a consistent behavior [38]. Our emotions determine behavior much in the same way as the emergence conditions described elsewhere [3].

After user input, we update emotional state based on three input characteristics, the format of the input (e.g., command, request, query, inform, threat), lexical analysis of the input (for politeness, personalization, level or statement of concern, threat or insult, profanity), and semantic content of the input (i.e., the interview topic).

We also rely on the very little research on emotion expression in schizophrenia and depression [4,27]. What we do understand is that schizophrenic facial expressions are least expressive but more responsive than depressed to positive stimuli, schizophrenics show more negative emotions than normals, schizophrenics are less likely than normals to demonstrate "felt" HAPPINESS, and that, when appropriate, normal individuals show HAPPINESS more than CONTEMPT, DISGUST, or ANGER, while schizophrenics are more likely to show CONTEMPT, but still are unlikely to show DISGUST or ANGER.

Gestural Models

For a system to train how to perceive behaviors indicative of mental illness, the behaviors must be realistic and responsive. Virtual human behavior takes many forms, including verbal output, gesture, body movement, and change of internal emotional and cognitive states. We worked with experts to devise algorithms and data structures that determine how the virtual human is to behave given inputs such as current emotional state and interpretation of user verbal input (see Table 1). For instance, using these Behavior Engine resources, the schizophrenic knows to begin pacing when his emotional state reaches a certain threshold of anger or fear, as it might when the user issues an unexpected command or when the topic of conversation is particularly upsetting.

Other gesture research also has influenced how we model virtual human behavior. For instance, children's gestures are larger than adults [6]. So are the gestures, as a rule, in some cultures compared to others [6]. Social roles and status (identity, age, familial relationship), emotion (e.g., sad look down), and truthfulness of an utterance all affect eye contact [38]. Unfamiliarity leads to placing oneself at bigger interactional distance than does familiarity, and avoiding aggressive gestures [18]. Finally, schizophrenics attempt to maintain greater distance, less involvement than healthy subjects from interaction partner, both on a mental and behavioral level [27].

CONCLUSION

Figure 6 shows some screen shots from JUST-TALK. We tested the next iteration of the application in the same course at NCJA in May 2002 and again in October.

Figure 6: JUST-TALK Screenshots

We have at least five technical development goals we feel we need to achieve to improve our virtual human architecture. First, we need to continue extensive testing and recording of inputs to improve recognition of student utterances. This is a key effort in making the virtual humans more realistic. The natural language processing methods provide many opportunities for tuning the system to provide better responses. This tuning is an iterative testing approach where the recognition accuracy can be measured and continually improved.

Second, we need to look to increasing visceral or intrinsic engagement, rather than just linguistic or conversational engagement. This will take two thrusts, adding realism to the background and replacing the generated speech with pre-recorded speech.

- The existing technology, upon which JUST-TALK is based, uses a modeled (i.e., not photo-realistic) VR background. To increase user engagement, a photo-realistic background created

using RTI's Video Reality technology will be integrated into the interaction environment. Integrating Video Reality will require interfacing to its rendering code from the Visualization Engine, and building a module to convey the camera geometry information associated with the video environment. When appropriate, the audio stream must be decoded and played synchronously with the image stream, while environmental sound must be provided separately. Branching and idle behaviors must be provided in the video stream, requiring careful filming and editing.

- We need to improve the realism of virtual subject's verbal replies. JUST-TALK used a computer-synthesized voice, but students and instructors indicated that using a recorded voice is highly desirable. Many nuances of determining psychological state can be picked up only through vocal inflections.

Third, we need to provide more gesture cues. In the newer, more immersive, visualization environment, the virtual human is allowed greater movement and gesture. We created hundreds of animations to portray as necessary, from pacing to sitting to fleeing, and from looking around to looking at the user to looking for nonexistent voices. We are using the gesture processing software in a variety of projects ranging from training emergency room staff to recognize potential bio-terrorism attacks to training Special Operations soldiers in first aid. JUST-TALK will import and leverage the gesture databases and software upgrades being developed by these parallel efforts. Still, work remaining to do includes providing more range of movement, in facial gestures and lip synching.

Fourth, we need to improve the emotional model. Working with experts, we need to reconsider the base emotional states and devised new methods for updating emotional state. Current emotional state now relies heavily on past emotional state, but also syntactic and semantic content of the user's input, personality, environmental cues, and time course. Still, we have ongoing work with psychiatrists, law officers, and other experts to develop more sophisticated emotional models, based on clinical experience and training of police crisis intervention teams.

Fifth, most students made some assessment about the subject in the training, noting that he was dressed relatively nicely, spoke as if he were well-educated and appeared to only recently have been having mental difficulties. But they noted that few other visual cues were available to them as students, including facial movements. And, although the subject reacted negatively to the police in several scenarios, students said he did not represent the extreme fear or dislike of police that students said they commonly encounter. The preset viewpoint (users could manipulate the view but rarely did) made it difficult to see the details of the subject's face, particularly when the subject stood behind the bench. We are now adjusting viewpoints appropriately to help provide more cues.

We are continuously refining our models. However, what we have already learned from fielding schizophrenic, depressed, and normal virtual humans we expect will lead to much more realistic, and engaging and effective, learning environments for interaction skills.

ACKNOWLEDGEMENTS

This material is based on work supported by RTI under SCDA R9898.001, the National Institute of Justice under Cooperative Agreement 2000-RD-CX-K002, and the National Science Foundation under Grant No. EIA-0121211. We thank Randy Dupont at the University of Tennessee at Memphis, Deborah Weisel at the North Carolina State University, and Martie Stanford and Pam Pope at the North Carolina Justice Academy for their assistance and efforts.

REFERENCES

1. André, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (2000). Exploiting Models of Personality and Emotions to Control the Behavior of Animated Interface Agents. In J. Rickel et al. (Eds.), Proceedings of the Fourth International Conference on Autonomous Agents (pp. 3-7), Barcelona, June 2000.
2. André, E., Rist, T., & Müller, J. (1999). Employing AI Methods to Control the Behavior of Animated Interface Agents. International Journal of Applied Artificial Intelligence, 13(4-5), 415-448.
3. Bécheiraz, P., & Thalmann, D. (1998). A Behavioral Animation System for Autonomous Actors Personified by Emotions. Proceedings of the First Workshop on Embodied Conversational Characters, Lake Tahoe, CA.
4. Berenbaum, H., & Oltmanns, T.F. (1992). Emotional Experience and Expression in Schizophrenia and Depression. Journal of Abnormal Psychology, 101(1), 37-44.
5. Cassell, J., Bickmore, T., Vilhjálmsson, H., & Yan, H. (2000). More Than Just a Pretty Face: Affordances of Embodiment. Proceedings of 2000 International Conference on Intelligent User Interfaces, New Orleans, LA.
6. Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Beckett, T., Douville, B., Prevost, S., & Stone, M. (1994). Animated Conversation: Rule-based Generation of Facial Display, Gesture and Spoken Intonation for Multiple Conversational Agents. Computer Graphics, 28(4), 413-420.
7. Cassell, J., & Thórisson, K.R. (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. International Journal of Applied Artificial Intelligence, 13(4-5), 519-538.
8. Cassell, J., & Vilhjálmsson, H.H. (1999). Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous. Autonomous Agents and Multi-Agent Systems, 2, 45-64.
9. Chovil, N. (1992). Discourse-Oriented Facial Displays in Conversation. Research on Language and Social Interaction, 25, 163-194.
10. Elliott, C. (1993). Using the Affective Reasoner to Support Social Simulations. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (pp. 194-200), Chambery, France, August 1993. Morgan Kaufmann.
11. Fehr, B., & Russell, J.A. (1984). Concept of Emotion Viewed From a Prototype Perspective. Journal of Experimental Psychology: General, 113(3), 464-486.
12. Frank, G.A., Guinn, C.I., Hubal, R.C., Stanford, M.A., Pope, P., & Lamm-Weisel, D. (2002). JUST-TALK: An Application of Responsive Virtual Human Technology. In Proceedings of I/ITSEC, Orlando, FL.
13. Funge, J., Tu, X., & Terzopoulos, D. (1999). Cognitive Modeling: Knowledge, Reasoning and Planning for Intelligent Characters. Computer Graphics Proceedings, ACM SIGGRAPH: 29-38.
14. Groves, R., & Couper, M. (1998). Nonresponse in Household Interview Surveys. New York: John Wiley & Sons, Inc.
15. Gunther-Mohr, C. (1998). Virtual Reality Training-Is It For You? Talk presented at the American Society for Training and Development International Conference, San Francisco, June 1998.
16. Guinn, C.I. (1996). Mechanisms for Mixed-Initiative Human-Computer Collaborative Discourse. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.

17. Guinn, C.I., & Montoya, R.J. (1998). Natural Language Processing in Virtual Reality. Modern Simulation & Training, 6, 44-45.
18. Guye-Vuillième, A., Capin, T.K., Pandzic, I.S., Magnenat Thalmann, N., & Thalmann, D. (1999) Non-verbal Communication Interface for Collaborative Virtual Environments. The Virtual Reality Journal, Springer.
19. Hubal, R.C., & Frank, G.A. (2001). Interactive Training Applications using Responsive Virtual Human Technology. Proceedings of I/ITSEC, Orlando, FL, November 2001.
20. Hubal, R.C., Frank, G.A., & Guinn, C.I. (2000). AVATALK Virtual Humans for Training with Computer Generated Forces. Proceedings of CGF-BR. Institute for Simulation & Training: Orlando FL.
21. Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., & West, S.L. (2000). The Virtual Standardized Patient-Simulated Patient-Practitioner Dialogue for Patient Interview Training. In J.D. Westwood, H.M. Hoffman, G.T. Mogel, R. A. Robb, & D. Stredney (Eds.), Medicine Meets Virtual Reality. IOS Press and Ohmsha, Amsterdam, 133-138.
22. Johnson, W.L., Rickel, J.W., & Lester, J.C. (2000). Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. International Journal of Artificial Intelligence in Education, 11, 47-78.
23. Kizakevich, P.N., McCartney, M.L., Nissman, D.B., Starko, K., & Smith, N.T. (1998). Virtual Medical Trainer: Patient Assessment and Trauma Care Simulator. Medicine Meets Virtual Reality-Art, Science, Technology: Healthcare (R)evolution, J.D. Westwood, H.M. Hoffman, D. Stredney, & S.J. Weghorst (Eds.), 309-315. IOS Press and Ohmsha: Amsterdam.
24. Klein, G. (1998). Sources of Power. Cambridge, MA: MIT Press.
25. Kline, C., & Blumberg, B. (1999). The Art and Science of Synthetic Character Design. Proceedings of the Symposium on AI and Creativity in Entertainment and Visual Art, Edinburgh.
26. Knapp, M.L., & Hall, J.A. (1997). Nonverbal Communication in Human Interaction (3rd ed.). Fort Worth: Harcourt Brace.
27. Krause, R., Steimer, E., Sänger-Alt, C., & Wagner, G. (1989). Facial Expression of Schizophrenic Patients and Their Interactions Partners. Psychiatry, 52, 1-11.
28. Lester, J.C., Stone, B.A., & Stelling, G.D. (1999). Lifelike Pedagogical Agents for Mixed-Initiative Problem Solving in Constructivist Learning Environments. User Modeling and User-Adapted Interaction, 9(1-2), 1-44.
29. Loyall, A.B., & Bates, J. (1997). Personality-Rich Believable Agents That Use Language. Proceedings of the First International Conference on Autonomous Agents, Marina del Rey CA.
30. Lundeberg, M., & Beskow, J. (1999). Developing a 3D-Agent for the August Dialogue System. AVSP Proceedings, Santa Cruz CA.
31. Moore, G. (2001). Talking Heads: Facial Animation in The Getaway. Gamasutra.
32. Olsen, D.E. (2001). The Simulation of a Human for Interpersonal Skill Training. In Proceedings of the 2001 Office of National Drug Control Policy (ONDCP) International Technology Symposium, June 25-28, 2001, San Diego CA.
33. Ortony, A., Clore, G. L., & Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge: Cambridge University Press.
34. Plutchik, R. (1997). The Circumplex as a General Model of the Structure of Emotions and Personality. In R. Plutchik & H.R. Conte (Eds.) Circumplex Models of Personality and Emotions (pp. 17-45). American Psychological Association.
35. Pollard, S., & Biermann, A. (2000). A Measure of Semantic Complexity for Natural Language Systems. Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems. ANLP-NAACL2000.
36. Prendinger, H., & Ishizuka, M. (2001). Social Role Awareness in Animated Agents. Fifth International Conference on Autonomous Agents (pp. 270-277), Montreal.
37. Ross, K.G., Pierce, L.G., Haltermann, J.A., & Ross, W.A. (1998). Preparing for the Instructional Technology Gap-A Constructivist Approach. Proceedings of I/ITSEC, Orlando FL.
38. Rousseau, D., & Hayes-Roth, B. (1997). Improvisational Synthetic Actors with Flexible Personalities. KSL Report #97-10, Stanford University.
39. Russell, J.A. (1997). How Shall an Emotion Be Called? In R. Plutchik & H.R. Conte (Eds.), Circumplex Models of Personality and Emotions (pp. 205-220). American Psychological Assn.
40. Shalif, I. (1991). The emotions and the dimensions of discrimination among them in daily life. Bar-Ilan University Ramat-Gan, Israel, unpublished dissertation.
41. Smith, R.W., & Gordon, S.A. (1997). Effects of Variable Initiative on Linguistic Behavior in Human-Computer Spoken Natural Language Dialog. Computational Linguistics, 23(1).
42. Wiggins, J.S., (Ed.) (1996). The Five Factor Model of Personality: Theoretical Perspectives. New York: The Guilford Press.